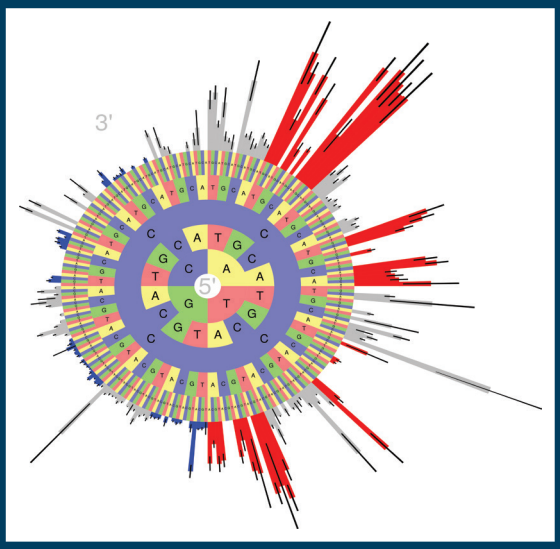
frontiers

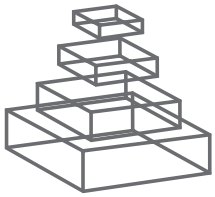
RESEARCH TOPICS



IMMUNE SYSTEM MODELING AND ANALYSIS

Topic Editors
Ramit Mehr, Miles Davenport,
Rob J. De Boer, Carmen Molina-Paris,
Michal Or-Guil and Veronika Zarnitsyna

frontiers in IMMUNOLOGY



frontiers

FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2015
Frontiers Media SA.
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88919-501-5
DOI 10.3389/978-2-88919-501-5

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

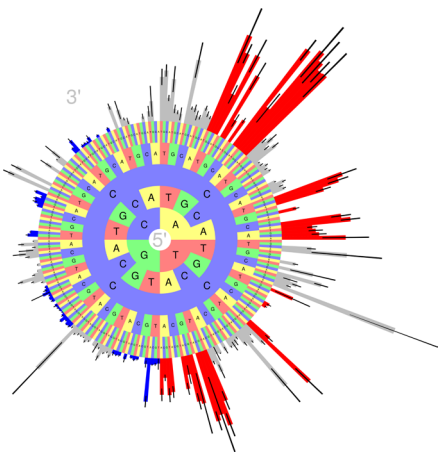
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

IMMUNE SYSTEM MODELING AND ANALYSIS

Topic Editors:
Ramit Mehr, Bar-Ilan University, Israel
Miles Davenport, University of New South Wales, Australia
Rob J. De Boer, Utrecht University, the Netherlands
Carmen Molina-Paris, University of Leeds, United Kingdom
Michal Or-Guil, Humboldt University Berlin, Germany
Veronika Zarnitsyna, Emory University, USA



The S5F model of somatic hypermutation targeting at the nucleotide C as a function of the surrounding nucleotides created by integration of multiple B cell repertoire sequencing data sets. Bar lengths indicate mutability as a function of the surrounding bases (specified on each ring), and bar colors indicate classic hot-spot WRC motifs (red) and cold-spot SYC motifs (blue). Copyright: Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JN, O'Connor KC, Hafler DA, Laserson U, Vigneault F, Kleinstein SH

The rapid development of new methods for immunological data collection – from multicolor flow cytometry, through single-cell imaging, to deep sequencing – presents us now, for the first time, with the ability to analyze and compare large amounts of immunological data in health, aging and disease. The exponential growth of these datasets, however, challenges the theoretical immunology community to develop methods for data organization and analysis. Furthermore, the need to test hypotheses regarding immune function, and generate predictions regarding the outcomes of medical interventions, necessitates the development of mathematical and computational models covering processes on multiple scales, from the genetic and molecular to the cellular and system scales.

The last few decades have seen the development of methods for presentation and analysis of clonal repertoires (those of T and B lymphocytes) and phenotypic (surface-marker based) repertoires of all lymphocyte types, and for modeling the intricate network of molecular and cellular interactions within the immune systems. This e-Book, which has first appeared as a 'Frontiers in Immunology' research topic, provides a comprehensive, online, open access snapshot of the current state of the art on immune system modeling and analysis.

Table of Contents

06 ***Immune system modeling and analysis***

Ramit Mehr

08 ***The structural basis of antibody-antigen recognition***

Inbal Sela-Culang, Vered Kunik and Yanay Ofran

21 ***Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin – implications for cross-reactivity of neutralizing antibodies***

Jing Sun, Ulrich J. Kudahl, Christian Simon, Zhiwei Cao, Ellis L. Reinherz and Vladimir Brusic

33 ***Pre-clustering of the B cell antigen receptor demonstrated by mathematically extended electron microscopy***

Gina J. Fiala, Daniel Kaschek, Britta Blumenthal, Michael Reth, Jens Timmer and Wolfgang W. A. Schamel

43 ***The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor***

Katherine J. L. Jackson, Marie J. Kidd, Yan Wang and Andrew M. Collins

55 ***Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data***

Gur Yaari, Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N. H. Stern, Kevin C. O'Connor, David A. Hafler, Uri Laserson, Francois Vigneault and Steven H. Kleinstein

66 ***Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures***

Gregory W. Schwartz and Uri Hershberg

73 ***Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions***

Gilad Liberman, Jennifer Benichou, Lea Tsaban, Jacob Glanville and Yoram Louzoun

83 ***Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation***

Thomas B. Kepler, Supriya Munshaw, Kevin Wiehe, Ruijun Zhang, Jae-Sung Yu, Christopher W. Woods, Thomas N. Denny, Georgia D. Tomaras, S. Munir Alam, M. Anthony Moody, Garnett Kelsoe, Hua-Xin Liao and Barton F. Haynes

93 ***A major hindrance in antibody affinity maturation investigation: we never succeeded in falsifying the hypothesis of single-step selection***

Michal Or-Guil and Jose Faro

97 ***A temporal model of human IgE and IgG antibody function***

Andrew M. Collins and Katherine J. L. Jackson

103 Self-tolerance in a minimal model of the idiotypic network

Robert Schulz, Benjamin Werner and Ulrich Behn

115 Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas

Miri Michaeli, Hilla Tabibian-Keissar, Ginette Schiby, Gitit Shahaf, Yishai Pickman, Lena Hazanov, Kinneret Rosenblatt, Deborah K. Dunn-Walters, Iris Barshack and Ramit Mehr

129 Addendum: Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas

Miri Michaeli, Hilla Tabibian-Keissar, Ginette Schiby, Gitit Shahaf, Yishai Pickman, Lena Hazanov, Kinneret Rosenblatt, Deborah K. Dunn-Walters, Iris Barshack and Ramit Mehr

132 Complementarity of binding motifs is a general property of HLA-A and HLA-B molecules and does not seem to effect HLA haplotype composition

Xiangyu Rao, Rob J. De Boer, Debbie van Baarle, Martin Maiers and Can Kesmir

138 Receptor pre-clustering and T cell responses: insights into molecular mechanisms

Mario Castro, Hisse M. van Santen, María Férez, Balbino Alarcón, Grant Lythe and Carmen Molina-París

149 Theories and quantification of thymic selection

Andrew J. Yates

164 From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection

Maria Sawicka, Gretta L. Stritesky, Joseph Reynolds, Niloufar Abourashchi, Grant Lythe, Carmen Molina-París and Kristin A. Hogquist

178 Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis

Gennady Bocharov, Tatyana Luzyanina, Jovana Cupovic and Burkhard Ludewig

185 Dynamical and mechanistic reconstructive approaches of T lymphocyte dynamics: using visual modeling languages to bridge the gap between immunologists, theoreticians, and programmers

Véronique Thomas-Vaslin, Adrien Six, Jean-Gabriel Ganascia and Hugues Bersini

191 A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children

Tharindi Hapuarachchi, Joanna Lewis and Robin E. Callard

197 Mathematical model of naive T cell division and survival IL-7 thresholds

Joseph Reynolds, Mark Coles, Grant Lythe and Carmen Molina-París

210 A mathematical model of immune activation with a unified self-nonself concept

Sahamoddin Khailaie, Fariba Bahrami, Mahyar Janahmadi, Pedro Milanez-Almeida, Jochen Huehn and Michael Meyer-Hermann

229 Harnessing the heterogeneity of T cell differentiation fate to fine-tune generation of effector and memory T cells

Chang Gong, Jennifer J. Linderman and Denise Kirschner

244 The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis

Adrien Six, Maria Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M. Walczak and Pierre Boudinot

260 Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling

Ilgar Z. Mamedov, Olga V. Britanova, Ivan V. Zvyagin, Maria A. Turchaninova, Dmitriy A. Bolotin, Ekaterina V. Putintseva, Yuriy B. Lebedev and Dmitriy M. Chudakov

270 Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire

Veronika I. Zarnitsyna, Brian D. Evavold, Louis N. Schoettle, Joseph N. Blattman and Rustom Antia

281 Huge overlap of individual TCR beta repertoires

Mikhail Shugay, Dmitriy A. Bolotin, Ekaterina V. Putintseva, Mikhail V. Pogorelyy, Ilgar Z. Mamedov and Dmitriy M. Chudakov

284 $CD4^+$ T cell-receptor repertoire diversity is compromised in the spleen but not in the bone marrow of aged mice due to private and sporadic clonal expansions

Eric Shifrut, Kuti Baruch, Hilah Gal, Wilfred Ndifon, Aleksandra Deczkowska, Michal Schwartz and Nir Friedman

294 Mother and child T cell receptor repertoires: deep profiling study

Ekaterina V. Putintseva, Olga V. Britanova, Dmitriy B. Staroverov, Ekaterina M. Merzlyak, Maria A. Turchaninova, Mikhail Shugay, Dmitriy A. Bolotin, Mikhail V. Pogorelyy, Ilgar Z. Mamedov, Vlasta Bobrynina, Mikhail Maschan, Yuri B. Lebedev and Dmitriy M. Chudakov

307 Mathematical models of the impact of IL2 modulation therapies on T cell dynamics

Kalet León, Karina García-Martínez and Tania Carmenate

328 Mechanisms underlying $CD4^+$ Treg immune regulation in the adult: from experiments to models

Marta Caridade, Luis Graca and Ruy M. Ribeiro

337 Mathematical modeling of oncogenesis control in mature T-cell populations

Sebastian Gerdes, Sebastian Newrzela, Ingmar Glauche, Dorothee von Laer, Martin-Leo Hansmann and Ingo Roeder

348 Inferring HIV escape rates from multi-locus genotype data

Taylor A. Kessinger, Alan S. Perelson and Richard A. Neher

361 Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus

Paola Carrillo-Bustamante, Can Keşmir and Rob J. de Boer

375 Corrigendum: Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus

Paola Carrillo-Bustamante, Can Keşmir and Rob J. De Boer

378 An interaction library for the FcεRI signaling network

Lily A. Chylek, David A. Holowka, Barbara A. Baird and William S. Hlavacek

394 Asymmetry in erythroid-myeloid differentiation switch and the role of timing in a binary cell-fate decision

Afnan Alagha and Alexey Zaikin

frontiers in
IMMUNOLOGY

EDITORIAL
published: 19 December 2014
doi: 10.3389/fimmu.2014.00644

Immune system modeling and analysis

Ramit Mehr*

Computational Immunology Lab, The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel
**Correspondence: ramit.mehr@biu.ac.il*

Edited and reviewed by:
Thomas L. Rothstein, The Feinstein Institute for Medical Research, USA

Keywords: immune system, mathematical modeling, lymphocytes, repertoire, immunomics

Immunologists currently face daunting challenges, as a result of the rapid development of new methods for immunological data collection, from high-throughput phenotyping to deep sequencing (1). These and similar methods keep generating humongous amounts of immunological data, which in turn challenge the theoretical immunology community to develop methods for data organization and analysis and mathematical and computational modeling. These challenges and methods were discussed in recent workshops, for example the Lymphocyte Repertoire Workshop (Institute of Advanced Studies of the Hebrew University, Jerusalem, early 2012, organized by myself), and the International Seminar on Multi-Scale Physics of Lymphocyte Development (Max Planck Institute for the Physics of Complex Systems, Dresden, Summer 2012, organized by M. Or-Guil et al.).

At about the same time, the organizers mentioned above were approached by the Frontiers editorial staff with the idea for a "Frontiers in Immunology" research topic, which was to provide a comprehensive, online, open access snapshot of the current state of the art on immune system modeling and analysis. The research topic was launched, edited, and finalized with the kind help of co-editors Rob de Boer, Miles Davenport, Carmen Molina-Paris, Michal Or-Guil, and Veronika Zarnitsyna. It has been a success, with 35 papers accepted for publication, which attests to the timeliness of the topic.

The papers included in this Research Topic reflect many of the issues that theoretical immunologists are struggling with. Some of the papers address old questions – such as the targeting of somatic hypermutation (2) and the resulting diversity of B cell repertoires (3, 4), how clonal selection operates in germinal centers (5–8); or how the T cell compartment develops (9–11) and changes with aging (12). However, these papers offer new viewpoints, which emerged thanks to the immunological "data revolution", in particular next-generation sequencing of lymphocyte repertoires. Others address new methods of extracting (13–15) and analyzing (16–18) comprehensive T and B cell phenotype and repertoire data, and delineate some of the first insights gleaned from sequencing studies regarding how these repertoires emerge, evolve, and function (19–25). Natural killer cells (26), myeloid cells (27), and structural immunology (28–31) are also represented.

My thanks go to the above-mentioned co-editors, to the responsive and efficient Frontiers editorial staff, to all the authors who contributed papers, and to the reviewers whose work has made publication of all these papers possible.

REFERENCES

1. Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol Lett* (2012) **148**:11. doi:10.1016/j.imlet.2012.08.002
2. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JN, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* (2013) **4**:358. doi:10.3389/fimmu.2013.00358
3. Jackson KJ, Kidd MJ, Wang Y, Collins AM. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front Immunol* (2013) **4**:263. doi:10.3389/fimmu.2013.00263
4. Michaeli M, Tabibian-Keissar H, Schiby G, Shahaf G, Pickman Y, Hazanov L, et al. Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas. *Front Immunol* (2014) **5**:264. doi:10.3389/fimmu.2014.00264
5. Schwartz GW, Hershberg U. Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. *Front Immunol* (2013) **4**:357. doi:10.3389/fimmu.2013.00357
6. Liberman G, Benichou J, Tsaban L, Glanville J, Louzoun Y. Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions. *Front Immunol* (2013) **4**:274. doi:10.3389/fimmu.2013.00274
7. Kepler TB, Munshaw S, Wiehe K, Zhang R, Yu JS, Woods CW, et al. Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front Immunol* (2014) **5**:170. doi:10.3389/fimmu.2014.00170
8. Or-Guil M, Faro J. A major hindrance in antibody affinity maturation investigation: we never succeeded in falsifying the hypothesis of single-step selection. *Front Immunol* (2014) **5**:237. doi:10.3389/fimmu.2014.00237
9. Yates AJ. Theories and quantification of thymic selection. *Front Immunol* (2014) **5**:13. doi:10.3389/fimmu.2014.00013
10. Reynolds J, Coles M, Lythe G, Molina-París C. Mathematical model of naive T cell division and survival IL-7 thresholds. *Front Immunol* (2013) **4**:434. doi:10.3389/fimmu.2013.00434
11. Hapuarachchi T, Lewis J, Callard RE. A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. *Front Immunol* (2013) **4**:366. doi:10.3389/fimmu.2013.00366
12. Shifrut E, Baruch K, Gal H, Ndifon W, Deczkowska A, Schwartz M, et al. CD4+ T cell-receptor repertoire diversity is compromised in the spleen but not in the bone marrow of aged mice due to private and sporadic clonal expansions. *Front Immunol* (2013) **4**:379. doi:10.3389/fimmu.2013.00379
13. Fiala GJ, Kaschek D, Blumenthal B, Reth M, Timmer J, Schamel WW. Preclustering of the B cell antigen receptor demonstrated by mathematically extended electron microscopy. *Front Immunol* (2013) **4**:427. doi:10.3389/fimmu.2013.00427
14. Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, et al. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol* (2013) **4**:456. doi:10.3389/fimmu.2013.00456
15. Chylek LA, Holowka DA, Baird BA, Hlavacek WS. An interaction library for the FcεRI signaling network. *Front Immunol* (2014) **5**:172. doi:10.3389/fimmu.2014.00172
16. Bocharov G, Luzyanina T, Cupovic J, Ludewig B. Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis. *Front Immunol* (2013) **4**:264. doi:10.3389/fimmu.2013.00264

17. Thomas-Vaslin V, Six A, Ganascia JG, Bersini H. Dynamical and mechanistic reconstructive approaches of T lymphocyte dynamics: using visual modeling languages to bridge the gap between immunologists, theoreticians, and programmers. *Front Immunol* (2013) **4**:300. doi:10.3389/fimmu.2013.00300
18. Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front Immunol* (2013) **4**:485. doi:10.3389/fimmu.2013.00485
19. Collins AM, Jackson KJ. A temporal model of human IgE and IgG antibody function. *Front Immunol* (2013) **4**:235. doi:10.3389/fimmu.2013.00235
20. Gong C, Linderman JJ, Kirschner D. Harnessing the heterogeneity of T cell differentiation fate to fine-tune generation of effector and memory T cells. *Front Immunol* (2014) **5**:57. doi:10.3389/fimmu.2014.00057
21. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. *Front Immunol* (2013) **4**:413. doi:10.3389/fimmu.2013.00413
22. León K, García-Martínez K, Carmenate T. Mathematical models of the impact of IL2 modulation therapies on T cell dynamics. *Front Immunol* (2013) **4**:439. doi:10.3389/fimmu.2013.00439
23. Caridade M, Graca L, Ribeiro RM. Mechanisms underlying CD4+ Treg immune regulation in the adult: from experiments to models. *Front Immunol* (2013) **4**:378. doi:10.3389/fimmu.2013.00378
24. Gerdes S, Newrzela S, Glauche I, von Laer D, Hansmann ML, Roeder I. Mathematical modeling of oncogenesis control in mature T-cell populations. *Front Immunol* (2013) **4**:380. doi:10.3389/fimmu.2013.00380
25. Kessinger TA, Perelson AS, Neher RA. Inferring HIV escape rates from multilocus genotype data. *Front Immunol* (2013) **4**:252. doi:10.3389/fimmu.2013.00252
26. Carrillo-Bustamante P, Kesmir C, de Boer RJ. Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus. *Front Immunol* (2014) **5**:20. doi:10.3389/fimmu.2014.00020
27. Alagha A, Zaikin A. Asymmetry in erythroid-myeloid differentiation switch and the role of timing in a binary cell-fate decision. *Front Immunol* (2013) **4**:426. doi:10.3389/fimmu.2013.00426
28. Sela-Culang I, Kunik V, Ofran Y. The structural basis of antibody-antigen recognition. *Front Immunol* (2013) **4**:302. doi:10.3389/fimmu.2013.00302
29. Sun J, Kudahl UJ, Simon C, Cao Z, Reinherz EL, Brusic V. Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin – implications for cross-reactivity of neutralizing antibodies. *Front Immunol* (2014) **5**:38. doi:10.3389/fimmu.2014.00038
30. Rao X, De Boer RJ, van Baarle D, Maiers M, Kesmir C. Complementarity of binding motifs is a general property of HLA-A and HLA-B molecules and does not seem to effect HLA haplotype composition. *Front Immunol* (2013) **4**:374. doi:10.3389/fimmu.2013.00374
31. Castro M, van Santen HM, Férez M, Alarcón B, Lythe G, Molina-París C. Receptor pre-clustering and T cell responses: insights into molecular mechanisms. *Front Immunol* (2014) **5**:132. doi:10.3389/fimmu.2014.00132

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 September 2014; accepted: 03 December 2014; published online: 19 December 2014.
*Citation: Mehr R (2014) Immune system modeling and analysis. Front. Immunol. **5**:644. doi: 10.3389/fimmu.2014.00644*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Mehr. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 08 October 2013
doi: 10.3389/fimmu.2013.00302

The structural basis of antibody-antigen recognition

Inbal Sela-Culang†, Vered Kunik† and Yanay Ofran*

The Goodman Faculty of Life Sciences, Bar Ilan University, Ramat Gan, Israel

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Gur Yaari, Yale University, USA
Chaim Putterman, Albert Einstein College of Medicine, USA

***Correspondence:**
Yanay Ofran, The Goodman Faculty of Life Sciences, Bar Ilan University, Ramat-Gan 52900, Israel
e-mail: yanay@ofranlab.org

† Inbal Sela-Culang and Vered Kunik have contributed equally to this work.

The function of antibodies (Abs) involves specific binding to antigens (Ags) and activation of other components of the immune system to fight pathogens. The six hypervariable loops within the variable domains of Abs, commonly termed complementarity determining regions (CDRs), are widely assumed to be responsible for Ag recognition, while the constant domains are believed to mediate effector activation. Recent studies and analyses of the growing number of available Ab structures, indicate that this clear functional separation between the two regions may be an oversimplification. Some positions within the CDRs have been shown to never participate in Ag binding and some off-CDRs residues often contribute critically to the interaction with the Ag. Moreover, there is now growing evidence for non-local and even allosteric effects in Ab-Ag interaction in which Ag binding affects the constant region and vice versa. This review summarizes and discusses the structural basis of Ag recognition, elaborating on the contribution of different structural determinants of the Ab to Ag binding and recognition. We discuss the CDRs, the different approaches for their identification and their relationship to the Ag interface. We also review what is currently known about the contribution of non-CDRs regions to Ag recognition, namely the framework regions (FRs) and the constant domains. The suggested mechanisms by which these regions contribute to Ag binding are discussed. On the Ag side of the interaction, we discuss attempts to predict B-cell epitopes and the suggested idea to incorporate Ab information into B-cell epitope prediction schemes. Beyond improving the understanding of immunity, characterization of the functional role of different parts of the Ab molecule may help in Ab engineering, design of CDR-derived peptides, and epitope prediction.

Keywords: antibody, CDRs, antigen, paratope, epitope, framework, constant domain

INTRODUCTION

Antibodies (Abs) have two distinct functions: one is to bind specifically to their target antigens (Ags); the other is to elicit an immune response against the bound Ag by recruiting other cells and molecules. The association between an Ab and an Ag involves myriad of non-covalent interactions between the epitope – the binding site on the Ag, and the paratopes – the binding site on the Ab. The ability of Abs to bind virtually any non-self surface with exquisite specificity and high affinity is not only the key to immunity but has also made Abs an enormously valuable tool in experimental biology, biomedical research, diagnostics and therapy. The diversity of their binding capabilities is particularly striking given the high structural similarity between all Abs. The availability of increasing amounts of structural data in recent years now allows for a much better understanding of the structural basis of Ab function in general, and of Ag recognition in particular. This review surveys the recent developments and the current gaps and challenges in this field. We focus specifically on the current understanding of the determinants within the Ab structure that contribute to Ag binding. We first discuss the motivations for, and applications of, the study of the structural basis of Ag recognition. Then we describe and discuss the Ab-Ag interface, with specific focus on the paratopes and the complementarity determining regions (CDRs), and their role in Ag binding. The last part focuses on the contribution of the non-CDRs parts of the Ab [i.e., framework regions (FRs) and the constant domains] to Ag binding and on the recent suggestions regarding non-local and allosteric effects in Ab function. Over the last few years numerous reviews have addressed issues that are related or tangential to the topics we review here. This includes reviews of the engineering of Abs (1), their stability (2), affinity maturation (3), and isotype selection (4). While these important topics are relevant to the findings and ideas we review here, they are beyond the scope of this review.

THE MOTIVATIONS FOR, AND APPLICATIONS OF, THE STUDY OF Ab-Ag RECOGNITION

UNDERSTANDING IMMUNITY AND AUTOIMMUNITY

The adaptive immune response involves two types of lymphocytes: T cells, which recognizes Ags that have been processed and their fragments are presented by MHC molecules, and B cells which produce soluble Abs that can identify also the intact Ag in its native form. While the way in which T cells recognize their epitopes has been extensively studied to a level that enables the successful prediction of T-cell epitopes (5, 6), the rules that govern Ab-Ag recognition, including which parts of the Ab structure underlie Ag recognition and how and why certain determinants on the Ag are selected as epitopes, are not as well characterized. Understanding the mechanisms that underlie Ab-Ag recognition, therefore, is crucial for understanding immunity.

The immune system enables Abs to distinguish between foreign and self molecules (7). Autoimmune diseases are characterized by the inappropriate response to self-Ags. It is not always clear what role is played by Abs and what role is played by other components of the immune system in autoimmunity. A variety of molecular mechanisms have been proposed, including sequestered Ags, molecular mimicry, and polyclonal B-cell activation (8). Better understanding of the underpinnings of Ab-Ag recognition may also shed light on these questions.

A MODEL FOR STUDYING BIO-MOLECULAR RECOGNITION

A fundamental characteristic of the immune system is its ability to continuously generate novel protein recognition sites. Ab-Ag interfaces, therefore, are often considered a model system for elucidating the principles governing biomolecular recognition (9–13). For example, Keskin (14) and McCoy et al. (15) used X-ray crystallographic structures of Ab-Ag complexes to elucidate principles of the molecular architecture of protein–protein interfaces. Other studies, however, view Ab-Ag interfaces as a specific case that may not allow for generalization to all types of protein–protein interfaces (16). Thus, large scale studies of protein–protein interactions often exclude Ab-Ag complexes from the dataset analyzed (16–19). It is, therefore, important to determine to what extent Ab-Ag complexes could serve as a general model for protein–protein interactions.

ANTIBODY ENGINEERING

The specificity of the Ab molecule to its cognate Ag has been exploited for the development of a variety of immunoassays, vaccinations, and therapeutics. Ab engineering may offer to expand the application of Abs by permitting improvements of affinity (20, 21) and specificity (22, 23). Understanding of the role each structural element in the Ab plays in Ag recognition is essential for successful engineering of better binders. The engineering of Abs is also important for the clinical use of Abs from non-human sources. Early studies on the use of rodent Abs in humans determined that they can be immunogenic (24). Humanization by grafting of the CDRs from a mouse Ab to a human FR is a commonly used engineering strategy for reducing immunogenicity (25, 26). In most cases, the successful design of high-affinity, CDR-grafted, Abs requires that key residues in the human acceptor FRs that are crucial for preserving the functional conformation of the CDRs will be back-mutated to the amino acids of the original murine Ab (26, 27). Several groups (28–30) used the experimentally determined 3-D structures of Ab-Ag complexes in the Protein Data Bank (PDB) (31) to determine which residues participate in Ag recognition and binding. Such knowledge can be exploited to identify residues that are important for the function of the Ab in general and for Ag recognition in particular and may guide Ab engineering (32, 33). Residues that help maintain the functional conformation of the CDRs, for example, can be used to improve Ab humanization efforts by CDR-grafting.

Ab EPITOPE PREDICTION

Antibody epitopes (sometimes referred to as B-cell epitopes) are the molecular structures within an Ag that make specific contacts with the Ab paratope. B-cell epitopes are used in the development of vaccines and in immunodiagnostics. Correct identification of B-cell epitopes within an antigenic protein, may open the door for the design of molecules (biologic or synthetic) that mimic potentially protective epitopes and could be used to raise specific Abs or be used as a prophylactic or therapeutic vaccines. Identification of B-cell epitopes could promote protective immunity in the context of emerging and re-emerging infectious diseases and potential bioterrorist threats. This may be achieved by choosing from among the putative epitopes those that may provide immunity (e.g., by eliciting Abs that hamper the molecular function of pathogenic Ags). The choice of such epitopes is believed to be relevant for understanding and controlling protective immunity. In the case of the vaccinia virus, for example, which was used as smallpox vaccine and is the only vaccine that has led to the complete eradication of an infectious disease from the human population, individuals possessing a high frequency of memory B-cells specific for major neutralizing Ags of the vaccinia virus are better protected from smallpox than individuals with a memory B-cell pool dominated by specificities for non-protective Ags (34). Thus, understanding the way in which an Ab recognizes its cognate epitope is of particular interest for vaccine design and disease prevention (35). Existing tools for identification of Ab epitopes (such as X-ray crystallography, pepscan, phage display, expressed fragments, partial proteolysis, mass spectrometry, and mutagenesis analysis) are not only expensive, laborious, and time consuming but also fail to identify many epitopes (36). When talking about protein Ags, most of these methods typically identify linear stretches as epitopes, while, arguably, most of the epitopes on protein Ags are conformational and even discontinuous. As for computational approaches, despite more than 30 years of efforts (37), existing B-cell epitope prediction methods are not accurate enough (38, 39) and are, therefore, not widely used. This is exemplified in **Figure 1**, in which the structure of hen egg lysozyme (HEL) Ag and three Abs that bind it are shown (**Figures 1A,B**), as well as the epitopes predicted by three different methods (**Figure 1C**).

In general, current methods are trying to identify epitopic residues based on the presence of features associated with residues that bind the Ab (40–50). One possible explanation for the failure of these methods is that the differences between epitopes and other residues are not substantial. Indeed, several analyses (51–53) have shown that the amino-acid composition of epitopes is essentially indistinguishable from that of other surface-exposed non-epitopic residues.

This lack of intrinsic properties that clearly differentiate between epitopic and non-epitopic residues and the fact (demonstrated in **Figure 1**) that most of the Ag surface may become a part of an epitope under some circumstances (54–57) suggest that epitopes depend, to a great extent, on the Abs that recognize them. This is exemplified in **Figure 1**: most of the HEL surface residues are part of an epitope of at least one Ab (**Figures 1A,B**), even though this figure shows only three Abs (out of dozens known to bind HEL). Almost all the residues predicted to be epitopic may be considered as correct predictions as they bind some Ab (**Figure 1C**) but also as false predictions as they don't bind the others. Similarly, predicting that a residue is not in an epitope may be either a true negative or a false negative, depending on the Ab considered. It has recently been suggested by us (Sela-Culang et al., submitted)

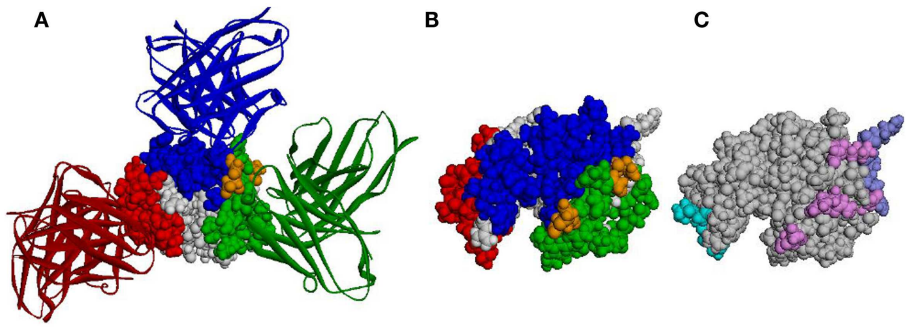


FIGURE 1 | Predicted epitopes vs. the actual epitopes of HEL. **(A)** The 3-D structure of HEL (CPK representation) together with three Abs (ribbon representation). PDB IDs 1JHL, 3D9A, and 1MLC were superimposed according to HEL structure. Epitope residues are colored blue, green, and red according to the corresponding Ab. Residues that are common to two epitopes are colored orange. **(B)** The structure of HEL colored according to the same three epitopes as in **(A)**, presented in a different orientation. **(C)** The structure of HEL colored according to the epitopes predicted by Discotope (light blue), ellipro (purple), and seppa (pink). Note, not all predicted residues of Discotope and ellipro are observable in the presented orientation.

and by others (58–60) that predicting epitopes should be done for a certain Ab. A similar concept was successfully applied in the case of T-cell epitope prediction methods: these methods do not examine the Ag for general features. Rather, different predictions are made, dependent on the specific MHC molecule binding and presenting the epitope to T cells.

THE ROLE OF CDRs AND THEIR DEFINITION

As shown in **Figure 2**, Abs are all-beta proteins consisting of four polypeptide chains: two identical heavy (H) chains and two identical light (L) chains (61). The light and heavy chains are linked by disulfide bonds to form the arms of a Y-shaped structure, each arm is known as a Fab (61). The Fab is composed of two variable domains (VH in the heavy chain and VL in the light chain) and two constant domains (CH1 and CL) (62). In the pairing of light and heavy chains, the two variable domains dimerize to form the Fv fragment which contains the Ag binding site. Within each variable domain lie six hypervariable loops (63), three in the light chain (L1, L2, and L3) and three in the heavy chain (H1, H2, and H3), supported by a conserved FR of β-sheets. The light and heavy variable domains fold in a manner that brings the hypervariable loops together to create the Ag binding site or paratope. Two additional domains of the heavy chain, CH2, and CH3, compose the Fc region which is responsible for mediating the biological activity of the Ab molecule.

CDRs IDENTIFICATION

As indicated by their names, CDRs are believed to account for the recognition of the Ag. Therefore, a major focus in analyzing the structural basis for Ag recognition has been in identifying the exact boundaries of the CDRs in a given Ab. It is a common practice to identify paratopes through the identification of CDRs. Kabat and co-authors (63, 64) were the first to introduce a systematic approach to identify CDRs in newly sequenced Abs. It was based on the assumption that CDRs are the most variable regions between Abs. Therefore, they aligned the (fairly limited) set of Ab sequences available at that time and identified the most variable positions. Based on the alignment, they introduced a numbering scheme for the residues in the hypervariable regions and determined which positions mark the beginning and the end of each CDR. As structural data became available, Chothia and Lesk (65, 66) manually analyzed a small number of experimentally solved 3-D structures and determined the structural location of the loop regions. The boundaries of the FRs and CDRs were determined and the latter have been shown to adopt a restricted set of conformations, based on the presence of certain residues at key positions in the CDRs and the flanking FRs. Their finding that Kabat's definitions of L1 and H1 are structurally incorrect led to the introduction of the Chothia numbering scheme. With the increase of available structural data, they ran their analysis anew and introduced a new definition of L1 (66) in 1989. In 1997 (67), however, they concluded that this correction was erroneous, and reverted to their original 1987 numbering scheme. While the Kabat and Chothia schemes treated separately the different families of immunoglobulin domains, Lefranc and colleagues (68, 69) proposed a unified numbering scheme (referred to as IMGT numbering scheme) for immunoglobulin variable domain genomic sequences, including Ab light and heavy variable domains, as well as T-cell receptor variable domains. To correlate between the sequence, structure, and domain folding behavior of all immunoglobulin variable domains, the Aho numbering scheme spatially aligned known 3-D structures of immunoglobulins and unified their numbering (70).

A drawback of the Kabat, Chothia, and IMGT numbering schemes is that CDRs length variability takes into account only the most common loop lengths; While both Kabat and Chothia schemes accommodate insertions with insertion letters (e.g., 27A), the IMGT scheme avoids the use of insertion codes for all but the least common very long loops, and the Aho numbering scheme places insertions and deletions symmetrically around a key position. However, Abs with unusually long insertions may be hard to annotate using these methods and, as a result, their CDRs may not be identified correctly. For instance, the recently determined 3-D crystal structure of two bovine Abs (71) reveal exceptionally long H3 CDRs (>60 residues), with long insertions which these methods cannot accommodate and thus cannot identify the CDRs of these Abs.

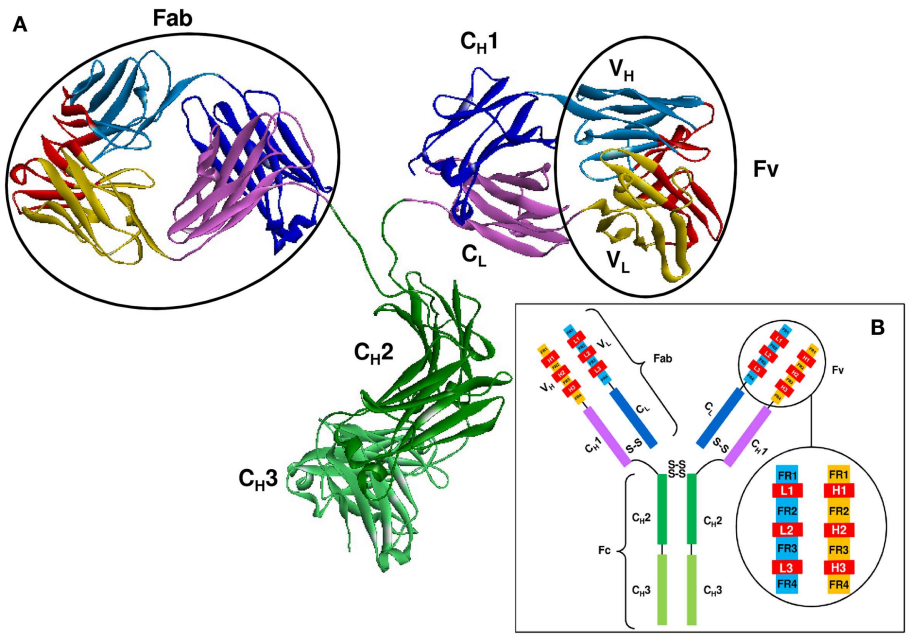


FIGURE 2 | The structure of an Ab molecule. **(A)** The 3-D structure of an Ab molecule (PDB ID: 1IGT). **(B)** A schematic representation of the Ab scaffold.

ARE CDRs GOOD PROXIES FOR THE PARATOPE?

While identification of paratopes is often done through identification of CDRs, not all the residues within the CDRs bind the Ag. In fact, an early analysis of the 3-D structures of Abs suggested that only 20–33% of the residues within the CDRs participate in Ag binding (72). In 1996, MacCallum and colleagues (73) performed a detailed residue-level analysis of Ag contacts. They suggested that contacting residues are more common at CDRs residues which are located at the center of the Ag combining site, and that non-contacting residues within the CDRs correspond with residues that are important for maintaining the structural conformations of the hypervariable loops and not necessarily for recognition of the Ag. Thus, they introduced a mapping of Ag-contacting propensities for each Ab position and proposed a new definition for CDRs based on these propensities. Padlan and co-workers (28) utilized Abs sequence and structure data to perform a by-position summary of Ag contacts. They found that the residues that are directly involved in the interaction with the Ag are also, in general, the most variable ones. They suggested that the residues that interact with the Ag should be called Specificity Determining Residues (SDRs).

The number of publicly available structures of Ab-Ag complexes increased in recent years to a level that enabled large-scale analyses. In a recent analysis (29) we utilized all available protein-Ab complexes in the PDB to identify the structural regions in which Ag binding actually occurs. This approach was implemented into a method dubbed Paratome (30, 74) that is based on a multiple structure alignment (MSTA) of all available Ab-Ag complexes in the PDB. The MSTA revealed regions of structural consensus where the pattern of structural positions that bind the Ag is highly similar among all Abs. These regions of structural binding consensus were termed antigen binding regions (ABRs). While CDRs, as identified by methods such as Kabat (63), Chothia (65), and IMGT (69), may miss ∼20% of the Ag binding residues, ABRs cover ∼96% of the residues that actually bind the Ag (30). To avoid confusions and cumbersome nomenclature, herein we generically refer to CDRs, SDRs, and ABRs as "CDRs" unless otherwise specified. **Figure 3** shows an example of CDRs as identified by Kabat, Chothia, IMGT, and Paratome for one Ab (anti-IL-15, PDB ID: 2XQB), compared to the actual Ag binding residues. It can be seen that in this example, some of the CDRs (e.g., L3, H3) identified by the four methods are almost identical, while in other CDRs (e.g., L2, H1, and H2) there are substantial differences between the methods. The MSTA of Abs with known 3-D structure also confirmed previous observations that there are structural positions within the CDRs in which none, or only a small percentage of the Abs contact the Ag. This is shown in **Figure 4** where an example of such a position is marked by a green arrow.

INTEGRALITY VS. MODULARITY

Designed systems are often characterized as either modular or integral. In a modular system different components, or modules, function independent of the function of other modules. The generation of Abs in the immune system is based on combining different elements, in a way that may be considered modular where each component is capable of binding the Ag regardless of the others. However, some analyses suggest that Ag binding warrants a more integrative view of the relationships between the different components of the Ab.

The binding-sites of interacting proteins are usually composed of surface patches that have good shape and electrostatic complementary (15, 75, 76). It has been shown that CDRs

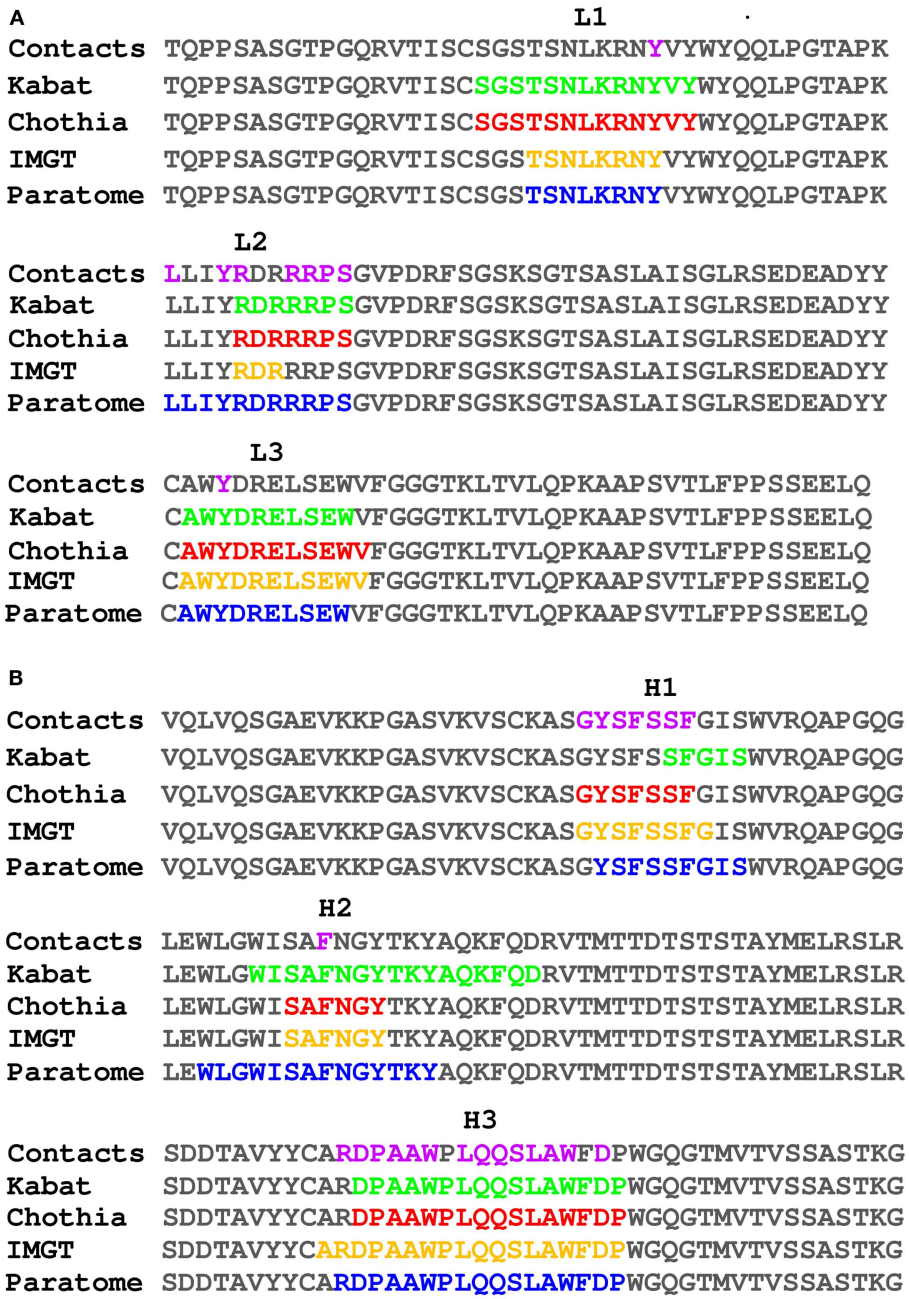


FIGURE 3 | Comparison of different CDR identification methods. The light **(A)** and heavy **(B)** chains of PDB ID 2XQB were numbered according to Kabat (colored green) and Chothia (colored red) using the Abnum tool (www.bioinf.org.uk/abs/abnum) and CDRs were extracted according to the CDR definitions table (www.bioinf.org.uk/abs/#cdrs). CDRs according to IMGT (colored orange) were identified using the IMGT-gap tool (www.imgt.org/3Dstructure-DB/cgi/DomainGapAlign.cgi). ABRs according to Paratome (colored blue) were identified using the Paratome server (www.ofranlab.org/paratome). Contacts (colored purple) between the Ab and IL-15 were defined using a 6-Å cutoff value.

are characterized by an amino-acid composition that is different from that of other protein loops (77) and also from other types of protein–protein interfaces (58). Thus, one would expect that epitopes, just like paratopes, should have a distinct amino-acid composition. However, several recent analyses (51, 53) have shown that this is not the case: while epitopes differ from other types of interfaces (10, 29, 60), their amino-acid composition is virtually the same as that of non-epitopic surface residues.

Several studies have shown that each CDR has its own unique amino-acid composition, different from the composition of the other CDRs (52, 58, 78). Additionally, we have shown that each CDR has a unique set of contact preferences, therefore, favoring

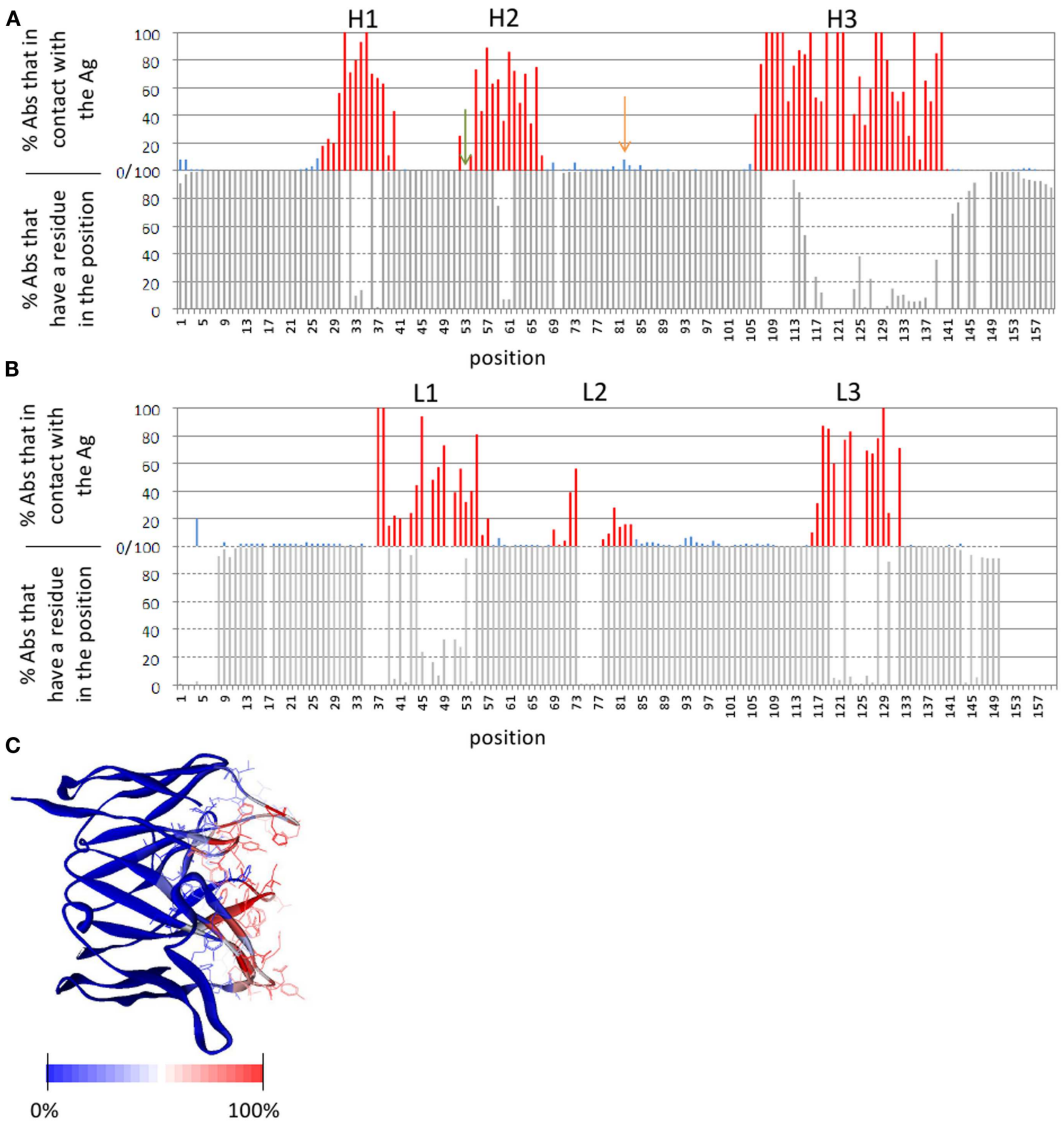


FIGURE 4 | Ab positions that contact the Ag. **(A,B)** The lower graphs show the percentage of Abs with known 3-D structure that have a residue in a given position (i.e., in other Abs there is a gap in the MSTA in that position). The upper graphs show the percentage of Abs that contact the Ag out of those Abs that have a residue in that position. **(A)** Depicts the heavy chain and **(B)** depicts the light chain. In the upper graphs, the ABRs are colored red and the FRs are colored blue. An example of a position within an ABR that is not in contact with the Ag in any of the Abs, is marked by a green arrow. An example of a position in the FRs that is in contact with the Ag in many (8%) of the Abs is marked by an orange arrow. **(C)** The Ab Fv domain (PDB ID: 1QFU) is colored according to the percentage of all Abs with known 3-D structure in which the residue in that position is in contact with the Ag: from red (100% of the Abs) to blue (0%). ABR residues are presented as lines. The definition of the ABRs is according to the Paratome server. A 6-Å cutoff value was used to define residues in contact. Percentages of contacts were calculated based on an MSTA of all protein Ab-Ag complexes in the PDB (30).

certain amino-acids over others (52). Dividing epitope residues into six subsets according to the CDR they bind, we found that each of the subsets has a distinct amino-acid composition, distinguishable from non-epitope surface (52). In other words, when the six subsets of epitope residues are considered together the unique composition of each subset disappears so that the overall amino-acid composition of the entire epitope is indistinguishable from the rest of the surface. Pathogenic epitopes may have evolved to resemble Ag surface to escape recognition. On the other hand, the integration of the six CDRs together, each with its own unique amino-acid composition and contact preferences, could be the evolutionary response of the immune system that enables Abs to recognize virtually any surface patch on the Ag.

Despite this integrated effect of the CDRs, Abs can be also considered as a modular system, composed of different elements (such as the Fab, VH and VL, or the six CDRs), which may bind the Ag on their own. Such smaller Ab fragments that retain Ag binding affinity and specificity, hold a great potential for drug design (79–81) as they have improved pharmacokinetics, tissue and tumor penetration, and can be produced more economically (80, 81).

They may also be combined with other fragments to yield better binders. Although such smaller fragments cannot induce effector function such as complement activation (due to the lack of the constant domains), they may neutralize the targeted Ag. Fab and single-chain variable (scFv) fragments usually maintain specific binding to the Ag (82). VH and VL fragments usually show sticky behavior, low solubility, and reduced Ag binding affinity (83–85), although, they sometime retain specificity to the Ag (83, 85–87).

The CDRs may provide additional level of modularity. According to the commonly accepted hotspot hypothesis, the binding energy of two proteins is largely determined by a very small number of critical interface residues (12, 88–90). Thus, one may wonder whether an individual CDR could bind the Ag on its own provided that it harbors hotspots. Several linear peptides containing one or more of the CDRs that retained Ag specificity have been reported (91–98). Although their affinity was usually in the micromolar range, it could be significantly improved by introducing relatively minor modifications (91, 99). However, many attempts to isolate and design such CDR derived peptides failed (100, 101). One possible reason is that a CDR, on its own, may not fold to the same conformation as in the context of the entire Fab, which may be crucial for binding. Cyclizing the CDR by adding Cys residues at its edges was suggested as a solution for this problem (96, 102–104). Another reason might lie in the fact that many attempts for the design of CDR-derived peptides are made based on CDR-H3, as it is considered to be the most important CDR for Ag binding (67, 105–107). However, the median length of ABR-H2 is substantially longer than that of H3, and both typically form the same number of interactions with the Ag (52). In addition, while ABR-H3 was shown to have the highest contribution to Ag binding energy on average (52), there are individual cases in which other CDRs are the dominant ones (52, 102). It is also possible that in some cases the binding depends on specific contacts from residues in different CDRs, which may preclude the design of CDR-derived peptides that maintain specificity. We have shown (102) that CDRs that are able to bind the Ag on their own have unique characteristics and, thus, can be computationally identified given the Ab-Ag complex structure. This may enhance the design of CDR-derived peptides that are not necessarily based on CDR-H3.

NON-CDR DETERMINANTS THAT HAVE A ROLE IN Ag BINDING

FR RESIDUES

Within the variable domain, the CDRs are believed to be responsible for Ag recognition, while the FR residues are considered a scaffold for the CDRs. However, it is now well established that some of the FR residues may play an important role in Ag binding (32, 108). As mentioned above, many such FR residues were identified during the process of Ab humanization by CDR grafting. While grafting only the CDRs usually results in a significant drop or a complete loss of binding, the binding affinity can be retained by back mutating some of the FR residues to the original murine sequence, emphasizing their role in Ag binding (26, 109–115).

Framework region residues that affect Ag binding can be divided into two categories. The first are FR residues that contact the Ag, thus are part of the binding-site (108, 109, 111, 116–123). Some of these residues are close in sequence to the CDRs (in fact they may be within the boundaries of CDRs according to some CDR identification methods, but not according to others, as shown in **Figure 3**). Other residues are those that are far from the CDRs in sequence, but are in close proximity to it in the 3-D structure. In particular, a loop in the heavy chain FR-3, sometimes referred to as CDR-H4, accounts for 1.3% of human Ab-Ag contacts (78, 124). This CDR-H4 is also enriched (in human Abs) in somatic hypermutations (Burkovitz et al., submitted). **Figure 4** shows positions that are not in the CDRs but are in contact with the Ag in many Abs [e.g., the one marked by an orange arrow (4A), which corresponds to CDR-H4].

In the second category of FR residues that affect Ag binding, are residues that are not in contact with the Ag, but affect Ag binding indirectly (108, 109, 120, 121). These residues can be further divided to those that are in spatial proximity to the CDRs, and those that are not. The former are assumed to affect binding by providing a structural support to the CDRs, enabling them to adopt the right conformation and orientation, shaping the binding-site required for Ag binding (32). For example, it has been suggested that a certain position in heavy chain FR-3, close in structure but not in sequence to CDR-H1 and CDR-H2, affects the orientation of CDR-H2 relative to CDR-H1 in such a way that a large side-chain packs between them and separates them while a small side-chain allows them to be closer to each other (109, 120). Nevertheless, this is not always true, as was shown in the case of the anti-lysozyme D1.3 Ab: while mutating Lys in this position to either Val, Ala, or Arg resulted in affinity difference, no structural change was observed (121).

Framework region residues that are more distant from the paratope are suggested to play a role in maintaining the overall structure of the Fv domains (32). However, these FR residues may also affect the Ag binding-site itself, by directing the relative orientation of the VH vs. the VL, and thus the orientation of the CDRs relative to each other (125–128). In particular, FR-2 residues were shown to play an important role in VH-VL interaction (129). Moreover, Masuda et al. (130) pointed to a specific position in the FR-2 loop, which controls the strength of the VH-VL interaction as well as its dependence on Ag binding. We have shown that the conformation of this loop changes upon Ag binding more than other residues in the FRs, and that the binding related conformational changes in this loop are similar in their magnitude to those of the CDRs (107). The potential role of the VH-VL interface in Ag binding is further supported by the observation that residues that are in the VH-VL interface (and are not a part of the Ab-Ag interface), are more likely to be mutated during the somatic hypermutation process, than residues that are not in either of these interfaces (Burkovitz et al., submitted).

Understanding the role of FR residues in Ag binding is crucial for efficient Ab design in general and for humanization in particular. Specifically, knowing in advance which FR residues may affect Ag binding, one may consider back-mutating these residues into their murine sequence, to improve affinity during CDR grafting. To this end, attempts were made to identify positions that contribute to Ag binding in multiple cases (32, 113, 119). For example, Haidar et al. (32) used a non-redundant dataset of Ab-Ag complex structures to identify positions that frequently contact the CDRs, and combined these positions with those that were back-mutated

frequently in the humanization literature. The 17 FR positions they identified were successfully used to design a combinatorial library for Ab humanization. Additional Abs, for which structures of both wild-type and a mutant(s) are available, may reveal the structural mechanisms by which each FR position affects Ag binding.

CONSTANT REGION

Until recently, Ab constant domains were considered responsible for the isotype and for effector function, such as complement activation, Fc receptor binding, avidity, and serum half-life (131). However, many studies now provide a strong evidence for a role for the constant region in Ag binding (131–147). There are many examples of Abs with identical variable domains but different isotypes that bind the same Ag with a different affinity or specificity (134–146). For instance, two Abs sharing identical variable domains but expressing different isotypes were shown to bind tubulin with significantly different affinities (135). Consistent with these studies, it has been shown that the complex of HEL and the Fv version of the HyHEL-10 Ab has an order of magnitude lower dissociation constant than the complex of HEL with the Fab version of this Ab (147). A probable explanation for this phenomenon would be an allosteric influence of the constant domains on the structure of the variable domains. Indeed, several structural studies provided some evidence for such structural effects (133, 146, 147). For example, Janda et al. (133) analyzed by Circular Dichroism (CD) spectra four different Ab isotypes of the 3E5 family that share identical variable domains, and showed that the different isotypes undergo different structural changes upon binding a common Ag. Similar results were obtained for anti-nuclear Abs as well: Xia et al. (146) compared four different isotypes of the PL9–11 anti-nuclear Ab sharing the same variable region, and found that the changes in secondary structure content (as revealed by CD analysis) as well as the wave length shifts of tryptophan fluorescence emission, upon Ag binding, are both isotype dependent. Recently, Tudor et al. (144) showed that this allosteric effect may control not only Ag binding affinity and specificity, but also the epitope recognized. They showed that two anti-HIV-1 IgG1 and IgA2 Abs with identical variable regions, recognize only partially overlapping epitopes.

Differences in affinity and specificity of Abs with the same variable region but different isotypes may play a role in autoimmunity if they occur in a self-reactive Ag. For example, different isotypes have been shown to be associated with different clinical outcomes for lupus erythematosus: a set of anti-PL9–11 Abs sharing the same variable domain but different isotypes were shown to bind DNA and chromatin, as well as the renal Ags, with different affinities that were associated with significant differences in renal pathogenicity *in vivo* and survival (148).

Several studies have suggested that allosteric effects in Abs may occur on the other direction as well: structural changes in the variable region caused by Ag binding may be transferred into the constant domains, potentially influencing effector activation and cellular response (131, 149–151). For example, Oda et al. (149) showed that the binding of staphylococcal protein A (SPA) or streptococcal protein G (SPG) to the constant region was inhibited by hapten binding in several Abs. A different example was provided by Horgan et al. (151) who observed differences in complement activation of two Abs which differ only in their VH domain.

An allosteric effect in Abs is further supported by a systematic computational analysis we have performed on all available free and Ag-bound pairs of structures (107). Many of the Ag-binding-related structural changes occur distant from the Ag binding-site, including changes in the relative orientation of the heavy and light chains in both the variable and constant domains as well as a change in the elbow angle between the variable and the constant domains. Moreover, the most consistent and substantial conformational change outside of the binding site was found in a loop in the heavy chain constant domain, which is a part of the CH1-CL interface, and is involved in complement binding (152).

What could be the mechanism for these allosteric effects? Changes in the constant domains sequence (different isotypes of the same Ab) or in its conformation (e.g., by effector binding) may lead to a rearrangement of the constant domains relative to each other and relative to the variable domains, which may result in a change to the VH-VL relative orientation (72), thus re-shaping the Ag binding-site (153–155).

The potential influence of the constant region on Ag affinity or specificity suggests that the process of class-switch may be considered, in combination with somatic hypermutations, as a mechanism for Ab diversity (131, 132). Engineering of an Ab of interest is usually associated with the optimization of its affinity to the Ag. Since the constant region may affect this affinity, the isotype selected should be carefully considered. Moreover, the constant region should be taken into account in vaccine design as well since different isotypes may bind the pathogenic Ag with different affinities, thus affecting the response to infection. For example, the anti HIV-1 IgG1 and IgA2 Abs mentioned above share the same variable region, nevertheless, they have been shown to block HIV-1 infection differently (144). While IgA2 blocked HIV-1 transcytosis and $CD4^+$ cell infection more efficiently, IgG1 and IgA2 act synergistically to block HIV-1 transfer from Langerhans cells to T cells. Thus, it has been suggested that a mucosal IgA-based vaccine response should complement an IgG-based vaccine response in blocking HIV-1 transmission.

CONCLUDING REMARKS

As Abs are one of the most versatile naturally occurring biosensors, it is of high importance to decipher the structural and molecular mechanisms by which they recognize and bind their Ags. Such knowledge is crucial for understanding immunity, may enable better prediction of Ab epitopes, and assist in Ab engineering.

While the commonly accepted view has been that CDRs hold the key for Ab-Ag recognition, recent findings indicate that not all the positions in the traditionally defined CDRs are important for binding. Furthermore, it has been shown that many positions that contribute critically to the binding energy reside outside of the transitional CDRs. Moreover, different CDR identification methods may often identify radically different stretches as "CDRs," indicating that CDRs are not well defined and thus are not necessarily a good proxy for the binding site. The hypervariable loops that accommodate the CDRs differ significantly from each other on various aspects. Understanding the way in which their binding preferences are integrated to yield the overall specificity of the Ab is an intriguing structural and biophysical challenge.

Accumulation of recent data suggests that elements that may be spatially distant from the Ag binding site also play a crucial role in Ag recognition. The unorthodox suggestion that non-local and even allosteric effects influence epitope recognition warrants additional analysis and research.

Addressing the open questions regarding the structural basis of Ag recognition requires additional structural data in the form of crystal structures of Abs bound to Ags of different types (proteins, peptides, nucleic acids, and haptens). Large-scale analysis of such structures will allow for the generation and testing of new hypotheses regarding the way in which Abs find and bind their epitopes.

ACKNOWLEDGMENTS

The authors are supported in part by NIAID contract N01-AI-90048C (http://www.niaid.nih.gov/).

REFERENCES

1. Vincent KJ, Zurini M. Current strategies in antibody engineering: Fc engineering and pH-dependent antigen binding, bispecific antibodies and antibody drug conjugates. *Biotechnol J* (2012) **7**:1444–50. doi:10.1002/biot.201200250
2. Wang W, Singh S, Zeng DL, King K, Nema S. Antibody structure, instability, and formulation. *J Pharm Sci* (2007) **96**:1–26. doi:10.1002/jps.20727
3. Sheedy C, MacKenzie CR, Hall JC. Isolation and affinity maturation of hapten-specific antibodies. *Biotechnol Adv* (2007) **25**:333–52. doi:10.1016/j.biotechadv.2007.02.003
4. Salfeld JG. Isotype selection in antibody engineering. *Nat Biotechnol* (2007) **25**:1369–72. doi:10.1038/nbt1207-1369
5. Brusic V, Bajic VB, Petrovsky N. Computational methods for prediction of T-cell epitopes – a framework for modelling, testing, and applications. *Methods* (2004) **34**:436–43. doi:10.1016/j.ymeth.2004.06.006
6. Kim Y, Sette A, Peters B. Applications for T-cell epitope queries and tools in the Immune Epitope Database and Analysis Resource. *J Immunol Methods* (2011) **374**:62–9. doi:10.1016/j.jim.2010.10.010
7. Frank SA. Specificity and cross-reactivity (Chapter 4). In: *Immunology and Evolution of Infectious Disease*. Princeton, NJ: Princeton University Press (2002). p. 33–54.
8. Rose NR, MacKay IR, editors. T-cells and autoimmunity. In: *The Autoimmune Diseases*, 4th ed. San Diego: Elsevier Academic Press (2006). 1160 p.
9. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* (1996) **93**:13–20. doi:10.1073/pnas.93.1.13
10. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* (1997) **272**:121–32. doi:10.1006/jmbi.1997.1234
11. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol* (1999) **285**:2177–98. doi:10.1006/jmbi.1998.2439
12. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol* (1998) **280**:1–9. doi:10.1006/jmbi.1998.1843
13. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins* (2002) **47**:334–43. doi:10.1002/prot.10085
14. Keskin O. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. *BMC Struct Biol* (2007) **7**:31. doi:10.1186/1472-6807-7-31
15. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. *J Mol Biol* (1997) **268**:570–84. doi:10.1006/jmbi.1997.0987
16. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* (2004) **338**:181–99. doi:10.1016/j.jmb.2004.02.040
17. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* (2003) **100**:5772–7. doi:10.1073/pnas.1030237100
18. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. *Proteins* (2005) **60**:353–66. doi:10.1002/prot.20433
19. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol* (2003) **325**:377–87. doi:10.1016/S0022-2836(02)01223-8
20. Marks JD, Griffiths AD, Malmqvist M, Clackson TP, Bye JM, Winter G. By-passing immunization: building high affinity human antibodies by chain shuffling. *Biotechnology (N Y)* (1992) **10**:779–83. doi:10.1038/nbt0792-779
21. Soderlind E, Ohlin M, Carlsson R. Complementarity-determining region (CDR) implantation: a theme of recombination. *Immunotechnology* (1999) **4**:279–85.
22. Hemminki A, Niemi S, Hautoniemi L, Soderlund H, Takkinen K. Fine tuning of an anti-testosterone antibody binding site by stepwise optimisation of the CDRs. *Immunotechnology* (1998) **4**:59–69. doi:10.1016/S1380-2933(98)00002-5
23. Ohlin M, Owman H, Mach M, Borrebaeck CA. Light chain shuffling of a high affinity antibody results in a drift in epitope recognition. *Mol Immunol* (1996) **33**:47–56. doi:10.1016/0161-5890(95)00123-9
24. Mirick GR, Bradt BM, Denardo SJ, Denardo GL. A review of human anti-globulin antibody (HAGA, HAMA, HACA, HAHA) responses to monoclonal antibodies. Not four letter words. *Q J Nucl Med Mol Imaging* (2004) **48**:251–7.
25. Jones PT, Dear PH, Foote J, Neuberger MS, Winter G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* (1986) **321**:522–5. doi:10.1038/321522a0
26. Queen C, Schneider WP, Selick HE, Payne PW, Landolfi NF, Duncan JF, et al. A humanized antibody that binds to the interleukin 2 receptor. *Proc Natl Acad Sci U S A* (1989) **86**:10029–33. doi:10.1073/pnas.86.24.10029
27. Co MS, Queen C. Humanized antibodies for therapy. *Nature* (1991) **351**:501–2. doi:10.1038/351501a0
28. Padlan EA, Abergel C, Tipper JP. Identification of specificity-determining residues in antibodies. *FASEB J* (1995) **9**:133–9.
29. Ofran Y, Schlessinger A, Rost B. Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J Immunol* (2008) **181**:6230–5.
30. Kunik V, Peters B, Ofran Y. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput Biol* (2012) **8**:e1002388. doi:10.1371/journal.pcbi.1002388
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* (2000) **28**:235–42. doi:10.1093/nar/28.1.235
32. Haidar JN, Yuan QA, Zeng L, Snavely M, Luna X, Zhang H, et al. A universal combinatorial design of antibody framework to graft distinct CDR sequences: a bioinformatics approach. *Proteins* (2012) **80**:896–912. doi:10.1002/prot.23246
33. Hanf KJ, Arndt JW, Chen LL, Jarpe M, Boriack-Sjodin PA, Li Y, et al. Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. *Methods* (2013). doi:10.1016/j.ymeth.2013.06.024
34. Amanna IJ, Slifka MK, Crotty S. Immunity and immunological memory following smallpox vaccination. *Immunol Rev* (2006) **211**:320–37. doi:10.1111/j.0105-2896.2006.00392.x
35. Yang XD, Yu XL. An introduction to epitope prediction methods and software. *Rev Med Virol* (2009) **19**:77–96. doi:10.1002/rmv.602
36. Xu XL, Sun J, Liu Q, Wang XJ, Xu TL, Zhu RX, et al. Evaluation of spatial epitope computational tools based on experimentally-confirmed dataset for protein antigens. *Chin Sci Bull* (2010) **55**:2169–74. doi:10.1007/s11434-010-3199-z
37. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A* (1981) **78**:3824–8. doi:10.1073/pnas.78.6.3824
38. Ponomarenko JV, Bourne PE. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct Biol* (2007) **7**:64. doi:10.1186/1472-6807-7-64
39. Blythe MJ, Flower DR. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci* (2005) **14**:246–8. doi:10.1110/ps.041059505

40. Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. *Immunome Res* (2010) **6**:6. doi:10.1186/1745-7580-6-6
41. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. *Nucleic Acids Res* (2005) **33**:W168–71. doi:10.1093/nar/gki460
42. Liang SD, Zheng DD, Zhang C, Zacharias M. Prediction of antigenic epitopes on protein surfaces by consensus scoring. *BMC Bioinformatics* (2009) **10**:302. doi:10.1186/1471-2105-10-302
43. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* (2008) **9**:514. doi:10.1186/1471-2105-9-514
44. Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B-cell epitopes. *Mol Immunol* (2009) **46**:840–7. doi:10.1016/j.molimm.2008.09.009
45. Rubinstein ND, Mayrose I, Martz E, Pupko T. Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinformatics* (2009) **10**:287. doi:10.1186/1471-2105-10-287
46. Sun J, Wu D, Xu T, Wang X, Xu X, Tao L, et al. SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Res* (2009) **37**:W612–6. doi:10.1093/nar/gkp417
47. Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics* (2008) **24**:1459–60. doi:10.1093/bioinformatics/btn199
48. Ambroise J, Giard J, Gala JL, Macq B. Identification of relevant properties for epitopes detection using a regression model. *IEEE/ACM Trans Comput Biol Bioinform* (2011) **8**:1700–7. doi:10.1109/TCBB.2011.77
49. Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC Bioinformatics* (2010) **11**:381. doi:10.1186/1471-2105-11-381
50. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. *BMC Bioinformatics* (2011) **12**:10. doi:10.1186/1471-2105-12-341
51. Janin J, Chothia C. The structure of protein-protein recognition sites. *J Biol Chem* (1990) **265**:16027–30.
52. Kunik V, Ofran Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng Des Sel* (2013). doi:10.1093/protein/gzt027
53. Kringelum JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol Immunol* (2013) **53**:24–34. doi:10.1016/j.molimm.2012.06.001
54. Benjamin DC, Berzofsky JA, East IJ, Gurd FR, Hannum C, Leach SJ, et al. The antigenic structure of proteins: a reappraisal. *Annu Rev Immunol* (1984) **2**:67–101. doi:10.1146/annurev.iy.02.040184.000435
55. Greenbaum JA, Andersen PH, Blythe M, Bui HH, Cachau RE, Crowe J, et al. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit* (2007) **20**:75–82. doi:10.1002/jmr.815
56. Novotný J, Handschumacher M, Haber E, Bruccoleri RE, Carlson WB, Fanning DW, et al. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc Natl Acad Sci U S A* (1986) **83**:226–30. doi:10.1073/pnas.83.2.226
57. Thornton JM, Edwards MS, Taylor WR, Barlow DJ. Location of "continuous" antigenic determinants in the protruding regions of proteins. *EMBO J* (1986) **5**:409–13.
58. Zhao L, Li JY. Mining for the antibody-antigen interacting associations that predict the B cell epitopes. *BMC Struct Biol* (2010) **10**(Suppl 1):S6. doi:10.1186/1472-6807-10-S1-S6
59. Zhao L, Wong L, Li JY. Antibody-specified b-cell epitope prediction in line with the principle of context-awareness. *IEEE/ACM Trans Comput Biol Bioinform* (2011) **8**:1483–94. doi:10.1109/TCBB.2011.49
60. Soga S, Kuroda D, Shirai H, Kobori M, Hirayama N. Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Eng Des Sel* (2010) **23**:441–8. doi:10.1093/protein/gzq014
61. Edelman GM, Benacerraf B. On structural and functional relations between antibodies and proteins of the gamma-system. *Proc Natl Acad Sci U S A* (1962) **48**:1035–42. doi:10.1073/pnas.48.6.1035
62. Putnam FW, Liu YS, Low TL. Primary structure of a human IgA1 immunoglobulin. IV. Streptococcal IgA1 protease, digestion, Fab and Fc fragments, and the complete amino acid sequence of the alpha 1 heavy chain. *J Biol Chem* (1979) **254**:2865–74.
63. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) **132**:211–50. doi:10.1084/jem.132.2.211
64. Kabat EA, Wu TT, Bilofsky H, Reid-Miller M, Perry H. *Sequence of Proteins of Immunological Interest*. Bethesda: National Institute of Health (1983).
65. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) **196**:901–17. doi:10.1016/0022-2836(87)90412-8
66. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* (1989) **342**:877–83. doi:10.1038/342877a0
67. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) **273**:927–48. doi:10.1006/jmbi.1997.1354
68. Giudicelli V, Chaume D, Bodmer J, Muller W, Busin C, Marsh S, et al. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* (1997) **25**:206–11. doi:10.1093/nar/25.1.206
69. Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* (2003) **27**:55–77. doi:10.1016/S0145-305X(02)00039-3
70. Honegger A, Pluckthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* (2001) **309**:657–70. doi:10.1006/jmbi.2001.4662
71. Wang F, Ekiert DC, Ahmad I, Yu W, Zhang Y, Bazirgan O, et al. Reshaping antibody diversity. *Cell* (2013) **153**:1379–93. doi:10.1016/j.cell.2013.04.049
72. Padlan EA. Anatomy of the antibody molecule. *Mol Immunol* (1994) **31**:169–217. doi:10.1016/0161-5890(94)90001-9
73. MacCallum RM, Martin AC, Thornton JM. Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol* (1996) **262**:732–45. doi:10.1006/jmbi.1996.0548
74. Kunik V, Ashkenazi S, Ofran Y. Paratome: an online tool for systematic identification of antigen binding regions in antibodies based on sequence or structure. *Nucleic Acids Res* (2012) **40**(Web Server issue):W521–4. doi:10.1093/nar/gks480
75. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* (1997) **272**:133–43. doi:10.1006/jmbi.1997.1234
76. Cohen GH, Silverton EW, Padlan EA, Dyda F, Wibbenmeyer JA, Willson RC, et al. Water molecules in the antibody-antigen interface of the structure of the Fab HyHEL-5-lysozyme complex at 1.7 A resolution: comparison with results from isothermal titration calorimetry. *Acta Crystallogr D Biol Crystallogr* (2005) **61**:628–33. doi:10.1107/S0907444905007870
77. Collis AV, Brouwer AP, Martin AC. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* (2003) **325**:337–54. doi:10.1016/S0022-2836(02)01222-6
78. Raghunathan G, Smart J, Williams J, Almagro J. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J Mol Recognit* (2012) **25**:103–13. doi:10.1002/jmr.2158
79. Nelson AL, Reichert JM. Development trends for therapeutic antibody fragments. *Nat Biotechnol* (2009) **27**:331–7. doi:10.1038/nbt0409-331
80. Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* (2005) **23**:1126–36. doi:10.1038/nbt1142

81. Hudson PJ, Souriau C. Engineered antibodies. *Nat Med* (2003) **9**:129–34. doi:10.1038/nm0103-129
82. Jain M, Kamal N, Batra SK. Engineering antibodies for clinical applications. *Trends Biotechnol* (2007) **25**:307–16. doi:10.1016/j.tibtech.2007.05.001
83. Ward ES, Güssow D, Griffiths AD, Jones PT, Winter G. Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli*. *Nature* (1989) **341**:544–6. doi:10.1038/341544a0
84. Rinfret A, Horne C, Dorrington KJ, Klein M. Noncovalent association of heavy and light chains of human immunoglobulins. IV. The roles of the CH1 and CL domains in idiotypic expression. *J Immunol* (1985) **135**:2574–81.
85. Berry MJ, Davies J. Use of antibody fragments in immunoaffinity chromatography. Comparison of FV fragments, VH fragments and paralog peptides. *J Chromatogr* (1992) **597**:239–45. doi:10.1016/0021-9673(92)80116-C
86. Pereira B, Benedict CR, Le A, Shapiro SS, Thiagarajan P. Cardiolipin binding a light chain from lupus-prone mice. *Biochemistry* (1998) **37**:1430–7. doi:10.1021/bi972277q
87. Dubnovitsky AP, Kravchuk ZI, Chumanevich AA, Cozzi A, Arosio P, Martsev SP. Expression, refolding, and ferritin-binding activity of the isolated VL-domain of monoclonal antibody F11. *Biochemistry (Mosc)* (2000) **65**:1011–8.
88. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science* (1995) **267**:383–6. doi:10.1126/science.7529940
89. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol* (2000) **10**:153–9. doi:10.1016/S0959-440X(00)00065-8
90. Ofran Y. *Prediction of Protein Interaction Sites. Computational Protein-Protein Interactions R Nussinov and G Schreiber*. Boca Raton: CRC Press (2009). p. 167–84.
91. Park BW, Zhang HT, Wu C, Berezov A, Zhang X, Dua R, et al. Rationally designed anti-HER2/neu peptide mimetic disables P185HER2/neu tyrosine kinases in vitro and in vivo. *Nat Biotechnol* (2000) **18**:194–8. doi:10.1038/72651
92. Polonelli L, Pontón J, Elguezabal N, Moragues MD, Casoli C, Pilotti E, et al. Antibody complementarity-determining regions (CDRs) can display differential antimicrobial, antiviral and antitumor activities. *PLoS One* (2008) **3**(6):e2371. doi:10.1371/journal.pone.0002371
93. Kang CY, Brunck TK, Kieber-Emmons T, Blalock JE, Kohler H. Inhibition of self-binding antibodies (autobodies) by a VH-derived peptide. *Science* (1988) **240**:1034–6. doi:10.1126/science.3368787
94. Taub R, Hsu JC, Garsky VM, Hill BL, Erlanger BF, Kohn LD. Peptide sequences from the hypervariable regions of two monoclonal anti-idiotypic antibodies against the thyrotropin (TSH) receptor are similar to TSH and inhibit TSH-increased cAMP production in FRTL-5 thyroid cells. *J Biol Chem* (1992) **267**:5977–84.
95. Saragovi HU, Fitzpatrick D, Raktabutr A, Nakanishi H, Kahn M, Greene MI. Design and synthesis of a mimetic from an antibody complementarity-determining region. *Science* (1991) **253**:792–5. doi:10.1126/science.1876837
96. Levi M, Sällberg M, Rudén U, Herlyn D, Maruyama H, Wigzell H, et al. A complementarity-determining region synthetic peptide acts as a miniantibody and neutralizes human immunodeficiency virus type 1 in vitro. *Proc Natl Acad Sci U S A* (1993) **90**:4374–8. doi:10.1073/pnas.90.10.4374
97. Tsumoto K, Misawa S, Ohba Y, Ueno T, Hayashi H, Kasai N, et al. Inhibition of hepatitis C virus NS3 protease by peptides derived from complementarity-determining regions (CDRs) of the monoclonal antibody 8D4: tolerance of a CDR peptide to conformational changes of a target. *FEBS Lett* (2002) **525**:77–82. doi:10.1016/S0014-5793(02)03090-9
98. Feng Y, Chung D, Garrard L, McEnroe G, Lim D, Scardina J, et al. Peptides derived from the complementarity-determining regions of anti-Mac-1 antibodies block intercellular adhesion molecule-1 interaction with Mac-1. *J Biol Chem* (1998) **273**:5625–30. doi:10.1074/jbc.273.10.5625
99. Feng J, Li Y, Zhang W, Shen B. Rational design of potent mimic peptide derived from monoclonal antibody: antibody mimic design. *Immunol Lett* (2005) **98**:311–6. doi:10.1016/j.imlet.2004.12.006
100. Lasonder E, Bloemhoff W, Welling GW. Interaction of lysozyme with synthetic anti-lysozyme D1.3 antibody fragments studied by affinity chromatography and surface plasmon resonance. *J Chromatogr A* (1994) **676**:91–8. doi:10.1016/0021-9673(94)00125-1
101. Schellekens GA. *Molecular Aspects of Antibody-Antigen Interactions: Size Reduction of a Herpes Simplex Virus Neutralizing Antibody and its Antigen*. Groningen: University of Groningen (1996).
102. Burkovitz A, Leiderman O, Sela-Culang I, Byk G, Ofran Y. Computational identification of antigen-binding antibody fragments. *J Immunol* (2013) **190**:2327–34. doi:10.4049/jimmunol.1200757
103. Bourgeois C, Bour JB, Aho LS, Pothier P. Prophylactic administration of a complementarity-determining region derived from a neutralizing monoclonal antibody is effective against respiratory syncytial virus infection in BALB/c mice. *J Virol* (1998) **72**:807–10.
104. Williams WV, Kieber-Emmons T, VonFeldt J, Greene MI, Weiner DB. Design of bioactive peptides based on antibody hypervariable region structures. Development of conformationally constrained and dimeric peptides with enhanced affinity. *J Biol Chem* (1991) **266**:5182–90.
105. Barrios Y, Jirholt P, Ohlin M. Length of the antibody heavy chain complementarity determining region 3 as a specificity-determining factor. *J Mol Recognit* (2004) **17**:332–8. doi:10.1002/jmr.679
106. Kuroda D, Shirai H, Kobori M, Nakamura H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* (2008) **73**:608–20. doi:10.1002/prot.22087
107. Sela-Culang I, Alon S, Ofran Y. A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J Immunol* (2012) **189**:4890–9. doi:10.4049/jimmunol.1201493
108. Sedrak P, Hsu K, Mohan C. Molecular signatures of anti-nuclear antibodies – contribution of heavy chain framework residues. *Mol Immunol* (2003) **40**:491–9. doi:10.1016/S0161-5890(03)00223-2
109. Xiang J, Sha Y, Jia Z, Prasad L, Delbaere L. Framework residue-71 and residue-93 of the chimeric B72.3 antibody are major determinants of the conformation of heavy-chain hypervariable loops. *J Mol Biol* (1995) **253**:385–90. doi:10.1006/jmbi.1995.0560
110. Verhoeyen M, Milstein C, Winter G. Reshaping human antibodies: grafting an antilysozyme activity. *Science* (1988) **239**:1534–6. doi:10.1126/science.2451287
111. Kettleborough CA, Saldanha J, Heath VJ, Morrison CJ, Bendig MM. Humanization of a mouse monoclonal antibody by CDR-grafting: the importance of framework residues on loop conformation. *Protein Eng* (1991) **4**:773–83. doi:10.1093/protein/4.7.773
112. Carter P, Presta L, Gorman CM, Ridgway JB, Henner D, Wong WL, et al. Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proc Natl Acad Sci U S A* (1992) **89**:4285–9. doi:10.1073/pnas.89.10.4285
113. Baca M, Presta L, O'Connor S, Wells J. Antibody humanization using monovalent phage display. *J Biol Chem* (1997) **272**:10678–84. doi:10.1074/jbc.272.16.10678
114. Rodríguez-Rodríguez ER, Ledezma-Candanoza LM, Contreras-Ferrat LG, Olamendi-Portugal T, Possani LD, Becerril B, et al. A single mutation in framework 2 of the heavy variable domain improves the properties of a diabody and a related single-chain antibody. *J Mol Biol* (2012) **423**:337–50. doi:10.1016/j.jmb.2012.07.007
115. Chiu WC, Lai YP, Chou MY. Humanization and characterization of an anti-human TNF-alpha murine monoclonal antibody. *PLoS One* (2011) **6**:e16373. doi:10.1371/journal.pone.0016373
116. Davies DR, Chacko S. Antibody structure. *Acc Chem Res* (1993) **26**:421–7. doi:10.1021/ar00032a005
117. Schroeder H, Hillson J, Perlmutter R. Structure and evolution of mammalian VH families. *Int Immunol* (1990) **2**:41–50. doi:10.1093/intimm/2.1.41
118. Amit A, Mariuzza R, Phillips S, Poljak R. 3-Dimensional structure of an antigen-antibody complex at 2.8 a resolution. *Science* (1986) **233**:747–53. doi:10.1126/science.2426778
119. Foote J, Winter G. Antibody framework residues affecting the

conformation of the hypervariable loops. *J Mol Biol* (1992) **224**:487–99. doi:10.1016/0022-2836(92)91010-M

120. Tramontano A, Chothia C, Lesk A. Framework residue-71 is a major determinant of the position and conformation of the 2nd hypervariable region in the VH domains of immunoglobulins. *J Mol Biol* (1990) **215**:175–82. doi:10.1016/S0022-2836(05)80102-0
121. Holmes M, Buss T, Foote J. Structural effects of framework mutations on a humanized anti-lysozyme antibody. *J Immunol* (2001) **167**:296–301.
122. Potter K, Hobby P, Klijn S, Stevenson F, Sutton B. Evidence for involvement of a hydrophobic patch in framework region 1 of human v4-34-encoded igs in recognition of the red blood cell I antigen. *J Immunol* (2002) **169**:3777–82.
123. Pospisil R, Youngcooper G, Mage R. Preferential expansion and survival of B-lymphocytes based on VH framework-1 and framework-3 expression: "positive" selection in appendix of normal and VH-mutant rabbits. *Proc Natl Acad Sci USA* (1995) **92**:6961–5. doi:10.1073/pnas.92.15.6961
124. Capra J, Kehoe J. Variable region sequences of 5 human immunoglobulin heavy chains of VH3 subgroup: definitive identification of four heavy chain hypervariable regions. *Proc Natl Acad Sci USA* (1974) **71**:845–8. doi:10.1073/pnas.71.3.845
125. Banfield M, King D, Mountain A, Brady R. V-L:V-H domain rotations in engineered antibodies: crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins* (1997) **29**:161–71. doi:10.1002/(SICI)1097-0134(199710)29:2<161::AID-PROT4>3.0.CO;2-G
126. Nakanishi T, Tsumoto K, Yokota A, Kondo H, Kumagai I. Critical contribution of VH-VL interaction to reshaping of an antibody: the case of humanization of anti-lysozyme antibody, HyHEL-10. *Protein Sci* (2008) **17**:261–70. doi:10.1110/ps.073156708
127. Stanfield R, Takimotokamimura M, Rini J, Profy A, Wilson I. Major antigen-induced domain rearrangements in an antibody. *Structure* (1993) **1**:83–93. doi:10.1016/0969-2126(93)90024-B
128. Tan P, Sandmaier B, Stayton P. Contributions of a highly conserved VHNL hydrogen bonding interaction to scFv folding stability and refolding efficiency. *Biophys J* (1998) **75**:1473–82. doi:10.1016/S0006-3495(98)74066-4
129. Essen L, Skerra A. The de-novo design of an antibody combining site – crystallographic analysis of the V-L domain confirms the structural model. *J Mol Biol* (1994) **238**:226–44. doi:10.1006/jmbi.1994.1284
130. Masuda K, Sakamoto K, Kojima M, Aburatani T, Ueda T, Ueda H. The role of interface framework residues in determining antibody V-H/V-L interaction strength and antigen-binding affinity. *Febs J* (2006) **273**:2184–94. doi:10.1111/j.1742-4658.2006.05232.x
131. Torres M, Casadevall A. The immunoglobulin constant region contributes to affinity and specificity. *Trends Immunol* (2008) **29**:91–7. doi:10.1016/j.it.2007.11.004
132. Casadevall A, Janda A. Immunoglobulin isotype influences affinity and specificity. *Proc Natl Acad Sci U S A* (2012) **109**:12272–3. doi:10.1073/pnas.1209750109
133. Janda A, Casadevall A. Circular dichroism reveals evidence of coupling between immunoglobulin constant and variable region secondary structure. *Mol Immunol* (2010) **47**:1421–5. doi:10.1016/j.molimm.2010.02.018
134. Dam TK, Torres M, Brewer CF, Casadevall A. Isothermal titration calorimetry reveals differential binding thermodynamics of variable region-identical antibodies differing in constant region for a univalent ligand. *J Biol Chem* (2008) **283**:31366–70. doi:10.1074/jbc.M806473200
135. Pritsch O, Hudry-Clergeon G, Buckle M, Petillot Y, Bouvet JP, Gagnon J, et al. Can immunoglobulin C(H)1 constant region domain modulate antigen binding affinity of antibodies? *J Clin Invest* (1996) **98**:2235–43. doi:10.1172/JCI119033
136. Pritsch O, Magnac C, Dumas G, Bouvet JP, Alzari P, Dighiero G. Can isotype switch modulate antigen-binding affinity and influence clonal selection? *Eur J Immunol* (2000) **30**(12):3387–95. doi:10.1002/1521-4141(2000012)30:12<3387::AID-IMMU3387>3.0.CO;2-K
137. Torres M, May R, Scharff MD, Casadevall A. Variable-region-identical antibodies differing in isotype demonstrate differences in fine specificity and idiotype. *J Immunol* (2005) **174**:2132–42.
138. Torres M, Fernández-Fuentes N, Fiser A, Casadevall A. The immunoglobulin heavy chain constant region affects kinetic and thermodynamic parameters of antibody variable region interactions with antigen. *J Biol Chem* (2007) **282**:13917–27. doi:10.1074/jbc.M700661200
139. McLean GR, Torres M, Elguezabal N, Nakouzi A, Casadevall A. Isotype can affect the fine specificity of an antibody for a polysaccharide antigen. *J Immunol* (2002) **169**:1379–86.
140. Cooper LJ, Shikhman AR, Glass DD, Kangisser D, Cunningham MW, Greenspan NS. Role of heavy chain constant domains in antibody-antigen interaction. Apparent specificity differences among streptococcal IgG antibodies expressing identical variable domains. *J Immunol* (1993) **150**:2231–42.
141. McCloskey N, Turner MW, Steffner P, Owens R, Goldblatt D. Human constant regions influence the antibody binding characteristics of mouse-human chimeric IgG subclasses. *Immunology* (1996) **88**:169–73. doi:10.1111/j.1365-2567.1996.tb00001.x
142. Michaelsen TE, Ihle Ø, Beckstrøm KJ, Herstad TK, Sandin RH, Kolberg J, et al. Binding properties and anti-bacterial activities of V-region identical, human IgG and IgM antibodies, against group B *Neisseria meningitidis*. *Biochem Soc Trans* (2003) **31**:1032–5. doi:10.1042/BST0311032
143. Liu F, Bergami PL, Duval M, Kuhrt D, Posner M, Cavacini L. Expression and functional activity of isotype and subclass switched human monoclonal antibody reactive with the base of the V3 loop of HIV-1 gp120. *AIDS Res Hum Retroviruses* (2003) **19**:597–607. doi:10.1089/088922203322230969
144. Tudor D, Yu H, Maupetit J, Drillet AS, Bouceba T, Schwartz-Cornil I, et al. Isotype modulates epitope specificity, affinity, and antiviral activities of anti-HIV-1 human broadly neutralizing 2F5 antibody. *Proc Natl Acad Sci U S A* (2012) **109**:12680–5. doi:10.1073/pnas.1200024109
145. Torosantucci A, Chiani P, Bromuro C, De Bernardis F, Palma AS, Liu Y, et al. Protection by anti-beta-glucan antibodies is associated with restricted beta-1,3 glucan binding specificity and inhibition of fungal growth and adherence. *PLoS One* (2009) **4**:e5392. doi:10.1371/journal.pone.0005392
146. Xia Y, Janda A, Eryilmaz E, Casadevall A, Putterman C. The constant region affects antigen binding of antibodies to DNA by altering secondary structure. *Mol Immunol* (2013) **56**:28–37. doi:10.1016/j.molimm.2013.04.004
147. Adachi M, Kurihara Y, Nojima H, Takeda-Shitaka M, Kamiya K, Umeyama H. Interaction between the antigen and antibody is controlled by the constant domains: normal mode dynamics of the HEL-HyHEL-10 complex. *Protein Sci* (2003) **12**:2125–31. doi:10.1110/ps.03100803
148. Xia Y, Pawar RD, Nakouzi AS, Herlitz L, Broder A, Liu K, et al. The constant region contributes to the antigenic specificity and renal pathogenicity of murine anti-DNA antibodies. *J Autoimmun* (2012) **39**:398–411. doi:10.1016/j.jaut.2012.06.005
149. Oda M, Kozono H, Morii H, Azuma T. Evidence of allosteric conformational changes in the antibody constant region upon antigen binding. *Int Immunol* (2003) **15**:417–26. doi:10.1093/intimm/dxg036
150. Piekarska B, Drozd A, Konieczny L, Król M, Jurkowski W, Roterman I, et al. The indirect generation of long-distance structural changes in antibodies upon their binding to antigen. *Chem Biol Drug Des* (2006) **68**:276–83. doi:10.1111/j.1747-0285.2006.00448.x
151. Horgan C, Brown K, Pincus SH. Effect of H-chain V-region on complement activation by immobilized immune-complexes. *J Immunol* (1992) **149**:127–35.
152. Vidarte L, Pastor C, Mas S, Blazquez AB, de los Rios V, Guerrero R, et al. Serine 132 is the C3 covalent attachment point on the CH1 domain of human IgG1. *J Biol Chem* (2001) **276**(41):38217–23.
153. Braden BC, Poljak RJ. Structural features of the reactions – between antibodies and protein antigens. *FASEB J* (1995) **9**:9–16.
154. Pellequer JL, Chen SW, Roberts VA, Tainer JA, Getzoff ED. Unraveling the effect of changes in conformation and compactness at the antibody V-L-V-H interface

upon antigen binding. *J Mol Recog* (1999) **12**:267–75. doi:10.1002/(SICI)1099-1352(199907/08)12:4<267::AID-JMR465>3.3.CO;2-0

155. Wilson IA, Stanfield RL. Antibody-antigen interactions – new structures and new conformational-changes. *Curr Opin Struct Biol* (1994) **4**:857–67. doi:10.1016/0959-440X(94)90267-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 August 2013; accepted: 12 September 2013; published online: 08 October 2013.

*Citation: Sela-Culang I, Kunik V and Ofran Y (2013) The structural basis of antibody-antigen recognition. Front. Immunol. **4**:302. doi: 10.3389/fimmu.2013.00302*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Sela-Culang, Kunik and Ofran. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 February 2014
doi: 10.3389/fimmu.2014.00038

Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin – implications for cross-reactivity of neutralizing antibodies

Jing Sun*[1,2], *Ulrich J. Kudahl*[1,3], *Christian Simon*[1,3], *Zhiwei Cao*[4], *Ellis L. Reinherz*[1,2,5] *and Vladimir Brusic*[1,2]

[1] *Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*
[2] *Department of Medicine, Harvard Medical School, Boston, MA, USA*
[3] *Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark*
[4] *School of Life Sciences and Technology, Tongji University, Shanghai, China*
[5] *Laboratory of Immunobiology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Juan C. Almagro, Pfizer, Inc., USA
Yanay Ofran, Bar-Ilan University, Israel

****Correspondence:***
Vladimir Brusic, Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, 77 Avenue Louis Pasteur, HIM 401, Boston, MA 02115, USA
e-mail: vladimir_brusic@dfci.harvard.edu

Influenza viruses continue to cause substantial morbidity and mortality worldwide. Fast gene mutation on surface proteins of influenza virus result in increasing resistance to current vaccines and available antiviral drugs. Broadly neutralizing antibodies (bnAbs) represent targets for prophylactic and therapeutic treatments of influenza. We performed a systematic bioinformatics study of cross-reactivity of neutralizing antibodies (nAbs) against influenza virus surface glycoprotein hemagglutinin (HA). This study utilized the available crystal structures of HA complexed with the antibodies for the analysis of tens of thousands of HA sequences. The detailed description of B-cell epitopes, measurement of epitope area similarity among different strains, and estimation of antibody neutralizing coverage provide insights into cross-reactivity status of existing nAbs against influenza virus. We have developed a method to assess the likely cross-reactivity potential of bnAbs for influenza strains, either newly emerged or existing. Our method catalogs influenza strains by a new concept named discontinuous peptide, and then provide assessment of cross-reactivity. Potentially cross-reactive strains are those that share 100% identity with experimentally verified neutralized strains. By cataloging influenza strains and their B-cell epitopes for known bnAbs, our method provides guidance for selection of representative strains for further experimental design. The knowledge of sequences, their B-cell epitopes, and differences between historical influenza strains, we enhance our preparedness and the ability to respond to the emerging pandemic threats.

Keywords: influenza virus, neutralizing antibodies, B-cell epitope, cross-reactivity, discontinuous peptide

INTRODUCTION

Influenza epidemics result in substantial morbidity and mortality (1). The World Health Organization (WHO) Global Influenza Network provides annual recommendations on antigenic variants to be included in the influenza vaccine formulations. Influenza virus has low-fidelity polymerases that result in high mutation rates (2). As a consequence, seasonal influenza viruses efficiently escape from acquired immunity in the human population through antigenic drift increasing the impact of seasonal influenza. The antigenic shift in influenza A viruses – the reassortment of multiple viral genomes resulting in new strains with recombined antigens – leads to occasional worldwide pandemics that result in significant morbidity and, usually, high mortality. High transmissibility of influenza combined with rapid mutation rates makes the discovery of novel influenza therapeutics an imperative (3). The main challenge in developing antibody-based prophylactics and therapeutic vaccine against influenza is to understand the variation generated by the virus and developing means to elicit broadly neutralizing antibody responses.

The majority of neutralizing antibodies (nAbs) generated during a normal immune response target hemagglutinin (HA) and block viral entry into host cells (4). However, significant sequence diversity among HA genes limits the protective breadth of these nAbs (5). This sequence diversity of influenza A virus is high – there are 17 HA serotypes that belong into two major groups called group 1 (Grp1: H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, and H17), and group 2 (Grp2: H3, H4, H7, H10, H14, and H15) (6). C179, the first neutralizing antibody reported to neutralize strains from H1 and H2 of influenza A virus, was isolated from mice immunized with the A/Okuda/57 (H2N2) strain (7). Later it was found that C179 was able to cross-neutralize H1, H2, H5, H6, and H9 subtypes (8–11). The next major advance in the field came about 15 years later (12), a novel class of human antibodies encoded by the V_H1–69 gene were discovered. Among these antibodies, a series of broadly neutralizing antibodies (bnAbs) have been described, such as CR6261 and F10 (13). Most bnAbs that neutralize influenza A virus have been reported to neutralize strains from either exclusively Grp1 or Grp2. FI6v3 (14) and 39.29 (5) are the only antibodies reported to neutralize human influenza isolates from both Grp1 and Grp2. Influenza B viruses are classified within a single influenza type, with two antigenically and genetically distinct lineages that co-circulate (15), represented

by the prototype viruses B/Victoria/2/1987 (Victoria lineage) and B/Yamagata/16/1988 (Yamagata lineage) (16). Antibody CR8071 (17) is a bnAb against influenza B viruses, with neutralizing ability for both Victoria and Yamagata lineages. bnAb CR9114 (17) binds a conserved epitope on the HA stem and was shown to neutralize all tested influenza A viruses. However, it did not show *in vitro* neutralizing activity against influenza B viruses at the tested concentrations (17).

Generally, the neutralizing effectiveness of these bnAbs was evaluated using representative strains from the subtypes of influenza A virus or lineages of influenza B virus. Because of the high variability of HA genes, such evaluation might result in a conclusion that is limited to the tested viral variants. To determine the landscape of nAbs and better understand their cross-reactivity properties, we performed a systematic study of B-cell epitopes of a selection of nAbs against influenza virus. Antibodies recognize discrete sites on the surface of macromolecule called B-cell epitopes (antigenic determinants). Some 10% of B-cell epitopes are linear peptides while 90% are formed from discontinuous amino acids that create surface patches through the three dimensional (3D) conformation of proteins (18). We defined a novel way of describing discontinuous motifs, using virtual peptides, to represent B-cell epitopes and further used this representation to estimate potential cross-reactivity and neutralizing coverage of these nAbs.

Functional characterization of the increasing number of nAbs and known crystal structures of these nAbs complexed with HA proteins enables us to precisely define their B-cell epitopes. A large number of sequences of influenza variants are available in public databases (19) enabling systematic bioinformatics analysis of cross-reactivity of nAbs against influenza virus. Such systematic analysis improves our understanding of antibody/antigen interactions, facilitates mapping of the known universe of target antigens, and allows the prediction of cross-reactivity. These methods and tools are useful for the design of broadly protective vaccines against emerging pathogens. This article describes a study of influenza HA cross-reactivity, but the method is applicable to any viral pathogen where information about nAbs and a collection of variant sequences of the target antigen are available.

MATERIALS AND METHODS

NEUTRALIZING ANTIBODIES AGAINST HEMAGGLUTININ

The names and specificities of nAb against influenza virus HA were collected from published papers. Twenty-two nAbs against influenza virus with crystal structures available in PDB were collected from published articles (**Table 1**). Fifteen of these nAbs target at the globular head of HA, and for the other seven, the binding sites are located on HA stem region.

Table 1 | Summary of well-characterized neutralizing antibodies against influenza virus.

Location	Neutralizing antibodies	PDB ID	Neutralizing breadth	Reference
Head	1F1	4GXU	H1	(20)
	2D1	3LZF	H1	(21)
	2G1	4HG4	H2	(17)
	8F8	4HF5	H2	(17)
	8M2	4HFU	H2	(17)
	BH151	*1EO8*	*A/X-31 (H3N2)*	(22)
	C05	4FQR	H1, H2, H3, H9	(23)
	CH65	3SM5	H1	(24)
	CH67	4HKX	H1	(25)
	CR8059	4FQK	Influenza B virus	(17)
	CR8071	4FQJ	Influenza B virus: Yamagata and Victoria	(17)
	HC19	*2VIR*	*A/X-31 (H3N2)*	(26)
	HC45	*1QFU*	*A/X-31 (H3N2)*	(27)
	HC63	*1KEN*	*A/X-31 (H3N2)*	(28)
	S139/1	4GMS	H1, H2, H3, H13, H16	(8, 29)
Stem	39.29	4KVN	H1, H2, H3	(5)
	C179	4HLZ	Grp1: H1, H2, H5, H6, H9	(30)
	CR6261	3GBN/3GBM	Grp1: H1, H2, H5, H9	(31, 32)
	CR8020	3SDY	Grp2: H3, H7, H10	(33)
	CR9114	4FQI/4FQV/4FQY	Grp1: H1, H2, H5, H6, H8, H9, H12 Grp2: H3, H4, H7, H10	(17)
	F10	3FKU	Grp1: H1, H2, H5, H6, H8, H9, H11	(13)
	FI6v3	3ZTJ/3ZTN	H1, H3, H5, H7	(14)

The nAbs in underlined italics are nAbs specific for strain A/X-31 (H3N2). The designation of two groups (Grp1 and Grp2) of influenza A virus subtypes are shown in ***Figure 1***.

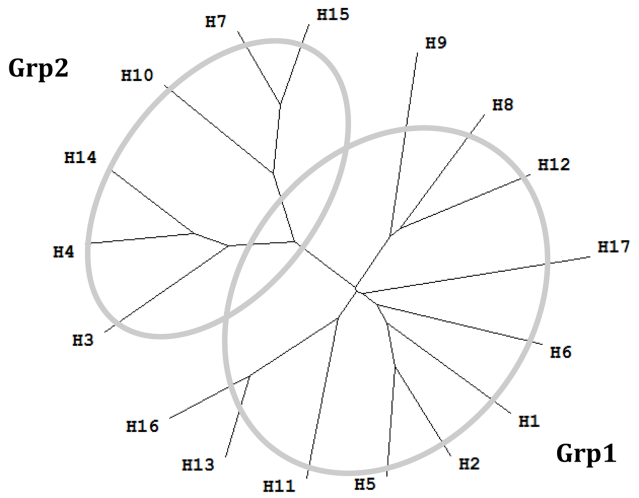


FIGURE 1 | Phylogenetic tree of 17 influenza A virus subtypes. Representative sequences were selected for each subtype (34) and the phylogenetic tree was made with ClustalX (35) and TreeView (36). H1: A/California/04/2009(H1N1); H2: A/Singapore/1/1957(H2N2); H3: A/Aichi/2/1968(H3N2); H4: A/duck/Czechoslovakia/1956(H4N6); H5: A/VietNam/1203/2004(H5N1); H6: A/chicken/California/431/2000(H6N2); H7: A/Turkey/Italy/8458/2002(H7N3); H8: A/Turkey/Ontario/6118/1968(H8N4); H9: A/Swine/Hong Kong/9/98(H9N2); H10: A/chicken/Germany/N/1949(H10N7); H11: A/duck/England/1/1956(H11N6); H12: A/duck/Alberta/60/1976(H12N5); H13: A/gull/Maryland/704/1977(H13N6); H14: A/Mallard/Astrakhan/263/1982(H14N5); H15: A/shearwater/West Australia/2576/79(H15N9); H16: A/black-headed gull/Sweden/2/99(H16N3); and H17: A/little yellow-shouldered bat/Guatemala/060/2010 (H17N10).

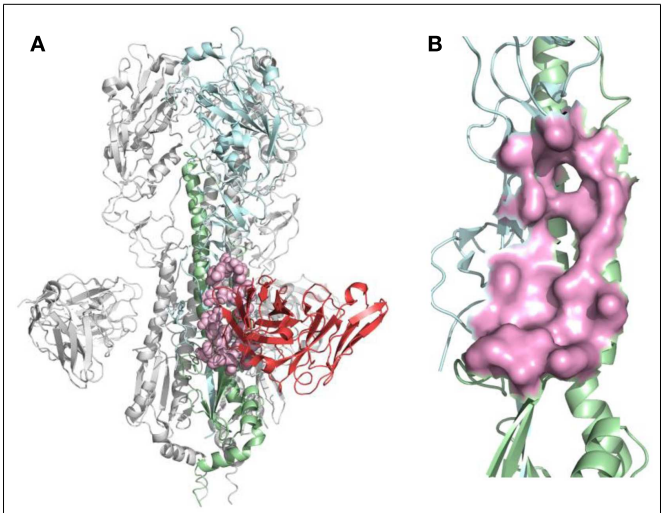


FIGURE 2 | B-cell epitope on the structure of neutralizing antibody F10 binding HA protein (PDB ID: 3FKU). **(A)** Complex of F10-HA [A/Vietnam/1203/04(H5N1)]. The structure is a HA trimer of three identical copies (one of them is colored as cyan and green; the other two are in gray). Each copy contains the HA1 (cyan) and HA2 (green) chain, also the heavy chain of F10 (red), the neutralized epitope is highlighted in pink; **(B)** Close-up view of neutralized epitope identified on the structure (highlighted as pink surface).

The majority of these nAbs were observed to bind or neutralize influenza A virus isolated either from Grp1 or Grp2. Antibodies FI6v3, CR9114, and 39.29 were shown to neutralize influenza strains within both Grp1 and Grp2 (5, 14, 27). Antibodies CR8059 and CR8071 (17) were the only two nAbs for influenza B virus. CR8059 is a light chain D95aN variant of CR8071. Since the mutation on CR8059 is not present at the binding interface and does not affect the binding, only CR8071 was used in the following study (17). The majority of these nAbs were shown to neutralize more than one strain, some of them are broadly neutralizing across subtypes of influenza A virus or lineages of influenza B virus. The Abs BH151, HC19, HC45, and HC63 were shown to specifically neutralize HA from the A/X-31(H3N2) strain. The available structures of nAb/HA complexes were downloaded from PDB (37).

VALIDATED INFLUENZA STRAINS BY NEUTRALIZING ANTIBODIES

Binding and neutralization assays were collected from published materials. Binding and non-binding strains were classified according to their affinity measurements. The thresholds used to discriminate binding and non-binding strains were inconsistent in different studies: the lowest affinity detectable values were set as 10^{-4} M (17), 10^{-5} M (33), and ~10^{-6} M (20). In some reports, nAbs showed positive binding results but did not display neutralization ability to the same strains [e.g., nAb CR9114 against strain B/Florida/4/2006 (Yamagata) (17)]. Because of the lack of standardized thresholds and ambiguous definition of binding, only results that indicate non-binding of antibodies were considered as useful information and were retained for the subsequent analysis as negatives.

Table 2 | B-cell epitope regions of the 22 neutralizing antibodies.

Binding site	**Influenza A virus**				**Influenza B virus**
	Sa site	**Near RBS**	**F subdomain**	**Stem base**	**Head base**
CROSS-REACTIVE NEUTRALIZING ANTIBODIES					
nAbs	**2D1**	**C05**	**CR6261**	**CR8020**	**CR8071**
		1F1	39.29		CR8059
		2G1	C179		
		8F8	CR9114		
		8M2	F10		
		CH65	FI6v3		
		CH67			
		S139/1			
Binding site	**Head base**	**RBS**	**Near RBS**		
X-31-SPECIFIC NEUTRALIZING ANTIBODIES					
nAbs	**BH151**	**HC19**	**HC63**		
	HC45				

The nAbs are classified as cross-reactive or X-31-specific. For each binding region, a representative nAb was selected (shown in bold) and its B-cell epitope was mapped on the structures shown in ***Figure 3****.*

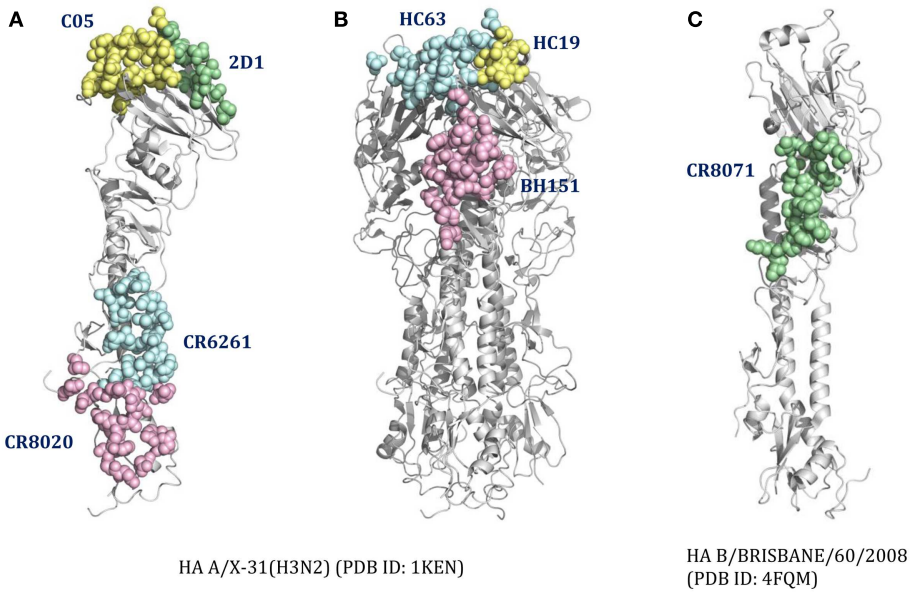


FIGURE 3 | The distinct B-cell epitope regions recognized by representative nAbs. The B-cell epitope regions of **(A)** represent cross-reactive nAbs against influenza A virus; **(B)** represent strain-specific nAbs against X-31(H3N2); **(C)** represent broadly nAb CR8071 against influenza B virus. The epitope regions of nAbs target influenza A virus were mapped on the monomer **(A)** or trimer **(B)** HA from A/X-31(H3N2) (PDB ID: 1KEN). The structure of B/Brisbane/60/2008 HA (PDB ID: 4FQM) was used as a template structure for influenza B virus. Different colors here were used for distinguishing B-cell epitope regions.

The neutralized and the escape strains were detected using the microneutralization assay (38) or HA inhibition assay (39). Several measurements were suggested in these studies:

1. The lowest concentration of nAb that displayed inhibition of hemagglutination or microneutralizing activity were set as either 2.5 μg/mL (40) or 5 μg/mL (41).
2. The 50% inhibitory concentration was set to $IC_{50} = 50\,\mu g/mL$ (17).
3. The effective concentration of antibody needed to inhibit at least 99% of viral infectivity was set as $EC_{99} = 100\,\mu g/mL$ (24, 25).

The HA sequences of strains that were experimentally validated for neutralization by studied antibodies ("validated strains") were retrieved from the literature. The influenza strains HA sequences were collected from the literature or, if absent, from the Influenza Knowledge Base (FLUKB)[1]. All experimentally validated strains were grouped into either neutralized strains or escape strains. The neutralized strains were selected based on reported experimental evidence. The escape strains included true escape strains as well as strains that were reported not to bind nAbs. We did not find any discrepancies in reported neutralizing properties across different studies used to collect functional data.

HEMAGGLUTININ SEQUENCES

All HA sequences were downloaded from the Influenza Knowledge Base (FLUKB[1], dated August 26th, 2013). After removing the incomplete sequences (fragments), 45,812 full-length HA sequences were left in the data set (HA sequence dataset) for further analysis.

GENERATION OF MULTIPLE SEQUENCE ALIGNMENT OF HEMAGGLUTININ SEQUENCES

The HA sequences of influenza strains from FLUKB were aligned using the MAFFT tool (42). The resulting multiple sequence alignment (MSA) results provided a consistent numbering scheme for all the further analyses. MSA were generated for both experimentally validated strains of HA and for all entries from FLUKB. For each nAb, every HA sequence from the crystal structure and from the experimentally validated strains were searched individually within the FLUKB database to find a strain with highest similarity using BLAST (43). This procedure was done to ensure that residue position mapping in following steps is consistent with the numbering scheme.

IDENTIFICATION OF B-CELL EPITOPES

B-cell epitope were identified from antigen–antibody structure, using a formula with the combination of the measurements of accessible surface area (ASA) and atom distance. For each residue from HA antigen, the ASA value was calculated using Naccess software (44) for both free HA and for HA coupled with an antibody. Residues r_i with ASA loss more than 20% were selected as epitope residues,

$$r_i \in \{\text{epitope residues}\} \text{ if } \frac{\text{ASA}_{\text{free}} - \text{ASA}_{\text{coupled}}}{\text{ASA}_{\text{free}}} > 0.2.$$

[1]http://research4.dfci.harvard.edu/cvc/flukb

Table 3 | B-cell epitope overlap for nAbs targeting HA head region.

	Sa	Near RBS									Sa	Near RBS								
	2D1	1F1	2G1	8F8	8M2	C05	CH65	CH67	S139/1		2D1	1F1	2G1	8F8	8M2	C05	CH65	CH67	S139/1
A98	−	+	−	+	+	+	+	+	+	A163	+	−	−	−	−	−	−	−	−
A125	+	−	−	−	−	−	−	−	−	A165	+	−	−	−	−	−	−	−	−
A126	+	−	−	−	−	−	−	−	−	A166	+	−	−	−	−	−	−	−	−
A128	+	−	+	−	−	−	−	−	−	A167	+	−	−	−	−	−	−	−	−
A130	+	−	+	+	+	−	−	+	−	A169	+	−	−	−	−	−	−	−	−
A131	−	−	−	−	−	+	−	−	+	A183	−	+	−	−	−	+	+	−	+
A132	−	−	+	+	+	−	−	−	−	A185	−	+	−	−	−	−	−	−	−
A133	−	+	+	+	+	+	−	−	−	A186	−	+	−	−	+	+	−	−	+
A134	−	−	+	+	+	+	+	+	+	A187	−	+	−	+	+	+	+	+	−
A135	−	+	−	−	−	+	+	+	+	A188	−	−	−	−	+	−	−	−	−
A136	−	−	+	+	+	+	+	+	+	A189	−	+	−	+	+	+	+	+	+
A137	−	−	+	+	+	+	+	+	+	A190	−	+	+	+	+	+	+	+	+
A140	−	−	−	+	−	−	−	−	−	A192	−	+	−	−	+	+	+	+	+
A143	−	−	−	+	−	−	−	−	−	A193	−	+	+	+	+	+	+	+	+
A144	−	−	−	+	−	−	−	−	−	A194	−	+	+	+	+	+	+	+	+
A145	−		+	+	+	+	−	−	+	A196	−	+	−	−	−	−	+	+	+
A153	−	+	+	+	+	+	+	+	+	A197	+	−	−	−	−	−	−	−	−
A155	−	+	+	+	+	+	+	+	+	A219	−	+	−	−	−	−	−	+	−
A156	−	+	+	+	+	+	+	+	+	A222	−	+	−	−	+	−	+	−	−
A157	+	−	+	−	−	−	−	−	+	A225	−	+	−	−	+	+	+	+	+
A158	+	−	+	+	+	+	+	+	+	A226	−	+	−	+	+	+	+	+	+
A159	+	+	+	+	+	+	+	+	+	A227	−	+	−	−	+	+	+	−	−
A160	+	−	−	−	−	−	+	+	+	A228	−	+	−	−	+	+	−	−	+
A161	+	−	−	−	−	−	−	−	−	A246	+	−	−	−	−	−	−	−	−
A162	+	−	−	−	−	−	−	−	−	A248	+	−	−	−	−	−	−	−	−

The epitope residue positions of nine nAbs were mapped to the 1EO8 structure chain A. The symbol "+" indicates a contact epitope residue by corresponding nAb, and the symbol "−" means it is not a epitope position. 2D1, with a different epitope area to other eight nAbs, is labeled in red.

The majority of contacts between two contacting atoms occur at distance smaller than 5 Å separation (45). Euclidean distance was calculated between atoms a_i and a_j using their coordinates $a_i(x_i, y_i, z_i)$ and $a_j(x_j, y_j, z_j)$ in PDB structure data,

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}.$$

Hemagglutinin residues r_i whose minimum atom distance to the closest nAb atom was within 4 Å were also incorporated in the epitope. The minimal atom distance was defined as:

$$d_{\min} = \min\{d_{ij}\},\ a_i \in \text{antigen residue } r_i, a_j \in \text{antibody residue } r_j,$$
$$r_i \in \{\text{epitope residue}\} \text{ if } d_{\min} < 4\text{Å}.$$

The residues that satisfy either of these two conditions (ASA loss or minimum distance) are considered to constitute a B-cell epitope.

The specific residues on HA that form hydrogen bonds, salt bridges, disulfide bonds, and covalent bonds between the HA and nAb were considered to define a B-cell epitope. The antigen/antibody interaction was further analyzed using PISA tool (46). The analysis of HA structures showed that all the hydrogen bonds, salt bridges, disulfide bonds, and covalent bonds between HA and nAb in each studied structure were incorporated in B-cell epitopes defined in the previous step.

EXTRACTION OF DISCONTINUOUS MOTIFS FROM VALIDATED STRAINS

For each nAb, using the MSA result and the standardized numbering, the residue positions of B-cell epitope identified from the HA/antibody crystal structure were mapped onto all HA sequence of validated strains. Then discontinuous motifs composed of mapped residues were extracted from these sequences. These discontinuous motifs were classified as either "neutralized" or "escape" motifs according to the experimental validation status of the corresponding strain.

MAPPING OF DISCONTINUOUS MOTIFS TO HA SEQUENCE DATASET

For each nAb, based on the MSA result, the residue positions of B-cell epitope identified from the HA/antibody crystal structure were mapped onto the HA sequence dataset. A "discontinuous peptide" composed of amino acids that form B-cell epitope, in order that they appear in the sequence, was extracted from each HA sequence. By comparing the discontinuous peptides to all validated neutralized and escape motifs from experimentally validated strains, each discontinuous peptide was classified as neutralized (if 100%

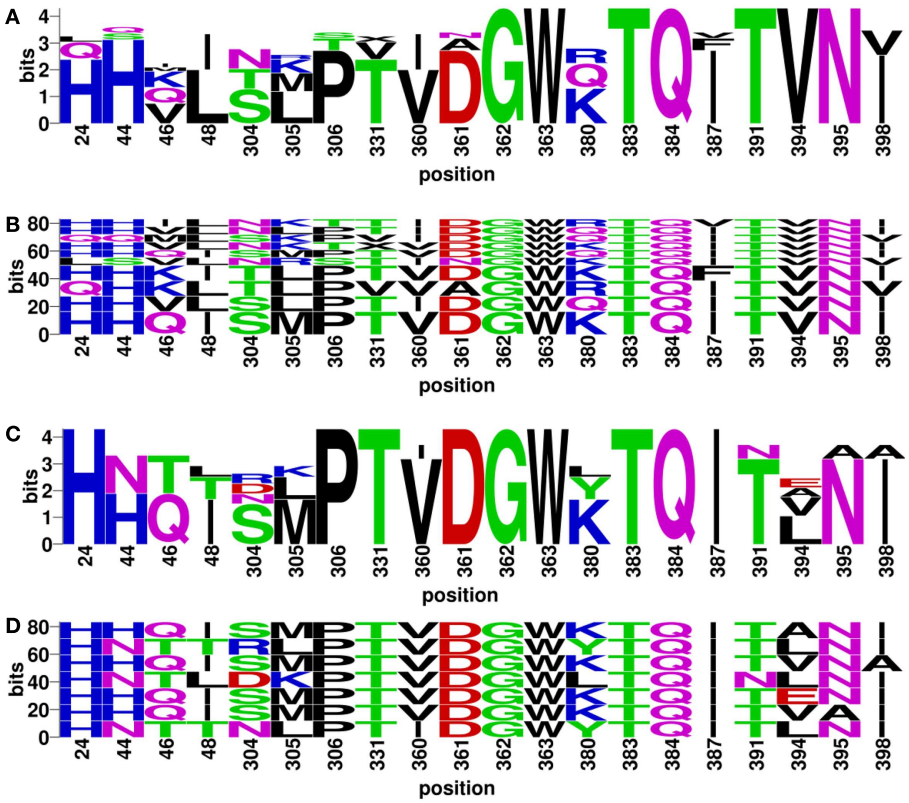


FIGURE 4 | Neutralized and escape discontinuous motifs from experimentally validated sequences with nAb F10. The WebLogo shows global **(A)** neutralized motifs, and **(C)** escape motifs and BlockLogo shows individual **(B)** neutralized motifs, and **(D)** escape motifs. The extracted discontinuous motif extracted from the structure (PDB ID: 3FKU, chain A and B), corresponds to the positions of reference sequence [FLU0293715, A/Viet Nam/1203/2004(H5N1)]: 24, 44, 46, 48, 304, 305, 306, 331, 360, 361, 362, 363, 380, 383, 384, 387, 391, 394, 395, and 398.

matching a neutralized epitope motif), escape (if 100% matching an escape epitope motif), or non-validated (if 100% matching validation data are missing). The term "discontinuous motif" indicates positions that define each B-cell epitope extracted from experimentally validated strains collected from publications, while term "discontinuous peptide" represents specific B-cell epitopes extracted from the HA sequence dataset.

RESULTS

B-CELL EPITOPE REGIONS

For each nAb, the B-cell epitope was identified from the crystal structure as described in Section "Materials and Methods." The structure of nAb F10-H5 (13) and identified epitope are illustrated in **Figure 2**. After B-cell epitopes of all studied nAbs were mapped to the same template structure, the overlapping of binding sites were found among different nAbs, particularly at the receptor-binding site (RBS), which is the necessary structure for binding to the sialic acid receptors during virus infection.

For cross-reactive nAbs against influenza A virus, four major binding locations on HA structure are apparent: two of them reside on the globular head of HA and the other two target the stem region of HA (**Table 2**; **Figure 3**). The RBS is a heavily targeted area, with overlapping epitopes defined by eight nAbs. The only nAb that binds HA head but not the RBS is 2D1 (21). The 2D1 recognizes the Sa site of A/South Carolina/1/1918(H1N1). Sa site is one of the earliest known antigenic sites (47), which is proximal to the receptor-binding pocket. The detailed comparison of epitope residue positions between 2D1 and the other HA head-targeted nAbs are listed in **Table 3**. In contrast to the Abs that interact with the HA head, a series of nAbs recognize another highly conserved helical region in the membrane-proximal HA stem. The epitopes on F subdomain (CR6261, 39.29, etc.) and stem base (CR8020) are adjacent to each other, with a small number of shared residues. The only broadly nAb neutralizing influenza B virus, CR8071 binds to the lower region of the globular head of HA – the "head base" (**Figure 3C**). All the remaining antibodies analyzed in our study bind specifically the HA on A/X-31(H3N2) strain. All X-31 specific nAbs complex with the membrane-distal domain of HA. NAbs BH151 and HC45 (22) recognize a single epitope located at the base of the eight-stranded antiparallel β-sheet structure. The HC19 binding site is adjacent to the RBS. The HC63 epitope shares several residues with HC19, thereby the antibody binding site overlaps the membrane-distal domains of two HA monomers.

EXPERIMENTALLY VALIDATED DISCONTINUOUS MOTIFS

Discontinuous motifs were extracted from the validated sequences as described in Section "Materials and Methods," and presented

by WebLogo (48) and BlockLogo[2] [Ref. (49)]. WebLogo figures consist of stacks of amino acids, while the overall height of the stack indicates the sequence conservation at that position, and the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. While BlockLogo is a web-based application for visualization of protein and nucleotide fragments, continuous protein sequence motifs, and discontinuous sequence motifs using calculation of block entropy from MSAs. The BlockLogo figures present the actual combinations of amino acids, and the height of each combination represents its relative frequency. In the nAb F10 as an example, the neutralized and escape discontinuous motifs are shown in **Figures 4A,C** (WebLogo figures), and **Figures 4B,D** (BlockLogo figures). WebLogos show a clear overall description of each residue conservation difference between individual neutralized and escape motifs. For example, 44N, 48T, 304R/D, 380L/Y, 391N, 394E/A/L on F10 epitope region are likely to contribute to the escape strains. In the BlockLogo figures, specific neutralized and escape B-cell epitopes of F10 were listed with their frequencies, which can be used for their direct comparison.

ANALYSIS OF VARIATION OF DISCONTINUOUS PEPTIDES IN HA SEQUENCES DATASET

For each nAb, the residue positions of their B-cell epitopes were mapped on the complete HA sequences dataset collected from the FLUKB. Amino acid strings representing discontinuous peptides were extracted from the HA sequence of each strain. The variability of discontinuous peptides and validated discontinuous motif coverage were analyzed for each nAb.

[2]http://research4.dfci.harvard.edu/cvc/blocklogo

For example, for the nAb F10, 589 different patterns of discontinuous peptides were generated among all 45,812 sequences in HA sequence dataset, using the F10 B-cell epitope identified from the crystal structure. In the next step, the discontinuous peptides were sorted according to their frequencies. The second most frequent peptide in FLUKB is identical an escape motif, while the 6th, 8th, and 19th are each identical to one of the neutralized motifs. However, the most frequent F10 discontinuous peptide in FLUKB (see text footnote 1) has not been experimentally tested (**Figure 5**), along with other 14 discontinuous peptides. The analysis of differences between the most frequent discontinuous peptide and neutralized or escape motifs was inconclusive. Therefore future experimental studies should include a representative sequence containing the discontinuous peptide HHVLSLPTVDGWLTQITVNI that is present in more than 10,000 entries in the FLUKB. We also recommend that motifs 1, 4, 5, 7, 9–18, and 20 are considered for the experimental validation. The remaining sequences are less common, each having <400 sequences in the data set.

The discontinuous peptides were generated and the variability was investigated for all cross-reactive nAbs (**Table 4**). The B-cell epitope regions on the HA stem are less variable as compared to the epitopes on the HA head. The specific result generated within each subtype in HA sequence dataset show similar patterns as for all subtypes (data not shown). This conclusion is consistent with our previous knowledge that the globular head of HA1 has a higher mutation rate than the stem (29), making the stem a more conserved region for bnAbs targeting.

DISCONTINUOUS MOTIFS COVERAGE IN HA SEQUENCES DATASET

The neutralized and escape discontinuous motifs of nAb F10 have covered 19 and 17% of FLUKB, respectively, while the discontinuous peptides from 64% of the strains have not been

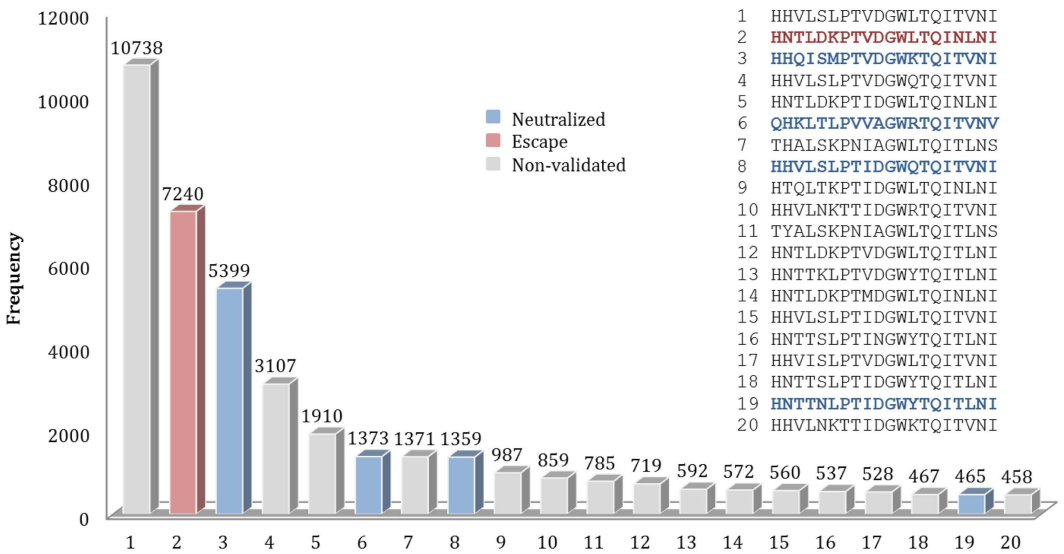


FIGURE 5 | Frequencies of top 20 discontinuous peptides (B-cell epitope of nAb F10) from the HA sequence dataset. FLU0243751 (A/Viet Nam/1203/2004) was used as reference HA sequence in the analysis of F10 B-cell epitopes. The corresponding positions of discontinuous peptides on FLU0243751 are: 24, 44, 46, 48, 304, 305, 306, 331, 360, 361, 362, 363, 380, 383, 384, 387, 391, 394, 395, and 398. Discontinuous peptides that were identical to neutralized motifs are shown in blue, while those identical to escape motifs are shown in red. The sequences of Top 20 most frequent discontinuous peptides are listed along with their validation status.

validated (**Figure 6A**). Viewed by the subtype, F10 neutralized coverage of subtypes H5, H8, H9, and H11 are higher (50–90%) than of H1 and H2 (5–20%), while the coverage of subtype H6 is negligible (8 in 1708 H6 sequences) (**Figure 6B**).

The motif coverage analysis within the 45,812 HA sequences was performed for all nAbs. For the nAbs with available cross-reactivity data, the motif coverages were different between the nAbs targeting the HA globular head and those targeting the stem part. The nAbs that bind stem normally have higher neutralized motif coverage than those that bind the globular head (**Figure 7**).

The motif coverage is shown as heat map for each subtype and each nAb (**Figure 8**). The nAbs (such as CR6261, CR9114, F10, and FI6v3) that target stem region are more cross-reactive – they cover more strains, and also more subtypes of influenza.

Table 4 | The number of different discontinuous peptides from B-cell epitopes of each nAb in the HA sequence dataset.

Neutralizing type	Neutralizing epitope regions		Neutralizing antibodies	Number of different discontinuous peptides
Influenza A virus				
	Head	Sa site	2D1	2,190
		Near RBS	1F1	2,887
			2G1	2,127
			8F8	2,885
			8M2	3,290
			C05	3,020
			CH65	2,727
			CH67	2,773
			S139/1	3,070
	Stem	F subdomain	39.29	983
			C179	755
			CR6261	658
			CR9114	663
			F10	589
			FI6v3	905
		Stem base	CR8020	620
Influenza B virus				
	Head base		CR8071	848

COMBINING OF NEUTRALIZING ANTIBODIES

For each sequence in the HA sequence dataset, 22 strings (discontinuous peptides) were extracted to represent 22 B-cell epitopes by all nAbs analyzed in this study. The majority (82.62%) of all strains in FLUKB have at least one discontinuous peptide that is identical to the validated neutralized motifs (**Table 5**). A small number (2.25%) of sequences can be neutralized by as many as seven nAbs.

Here, we propose a combination of nAbs, where a small number of nAbs can cover a large proportion of influenza strains. The nAbs FI6v3, F10, CR9114, and CR8071 (**Figure 9A**) were selected, and the neutralized coverage has increased from 18.91% (F10), 4.06% (CR8071), 43.89% (CR9114), and 58.44% (FI6v3) to 78.45% (**Figure 9B**) when these antibodies were combined. These nAbs also covered most subtypes of influenza A virus and both lineages in influenza B virus (**Figure 9C**).

DISCUSSION

This study presents an overview of binding specificities of reported nAbs, as well as an estimate of their neutralization and escape coverage (neutralization effectiveness) in more than 45,000 HA sequences available in FLUKB. The variety and frequency of discontinuous peptides within different B-cell epitopes have been analyzed in the HA data set. The results of the analysis of discontinuous peptides provide insights into further experimental design: strains with peptides that have high frequency among the strain populations should be given priority for experimental validation and their neutralizing status for specific nAbs.

Of note, additional sequence changes in HA outside the nAb epitope may result in either local or quaternary structural alterations that impacts antibody binding to the epitope *per se*.

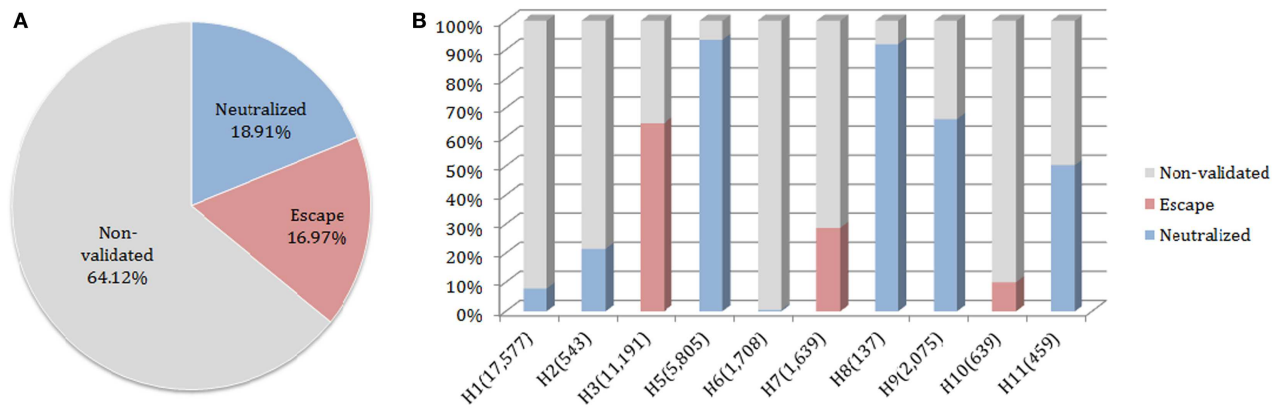


FIGURE 6 | Motif coverage for nAb F10. **(A)** The coverage of neutralized and escape discontinuous motifs, and non-validated discontinuous peptides within 45,812 HA sequences extracted from the FLUKB; **(B)** The motif coverage by subtype, the numbers in brackets indicate the number of sequence within the specific subtype (among 45,812 HA sequences).

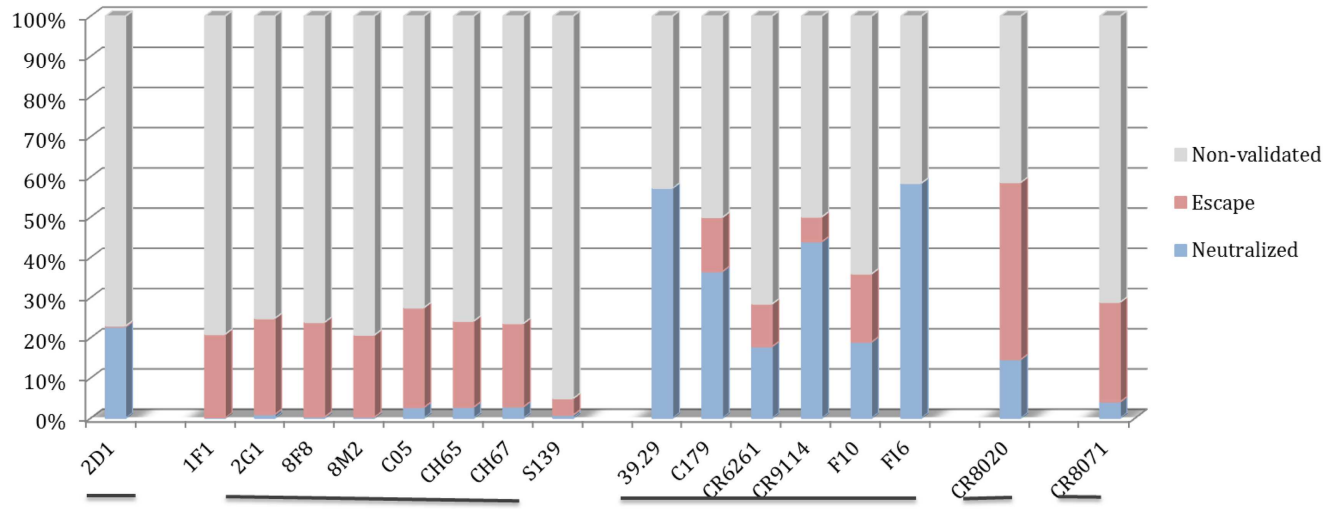


FIGURE 7 | Discontinuous motif coverage in the HA sequence dataset for cross-reactive nAbs. Only the coverage data for cross-reactive nAbs are shown here. The nAbs are grouped based on their binding locations and influenza types, from left to right: Sa site, near RBS, F subdomain, stem base on influenza A virus, and head base on influenza B virus.

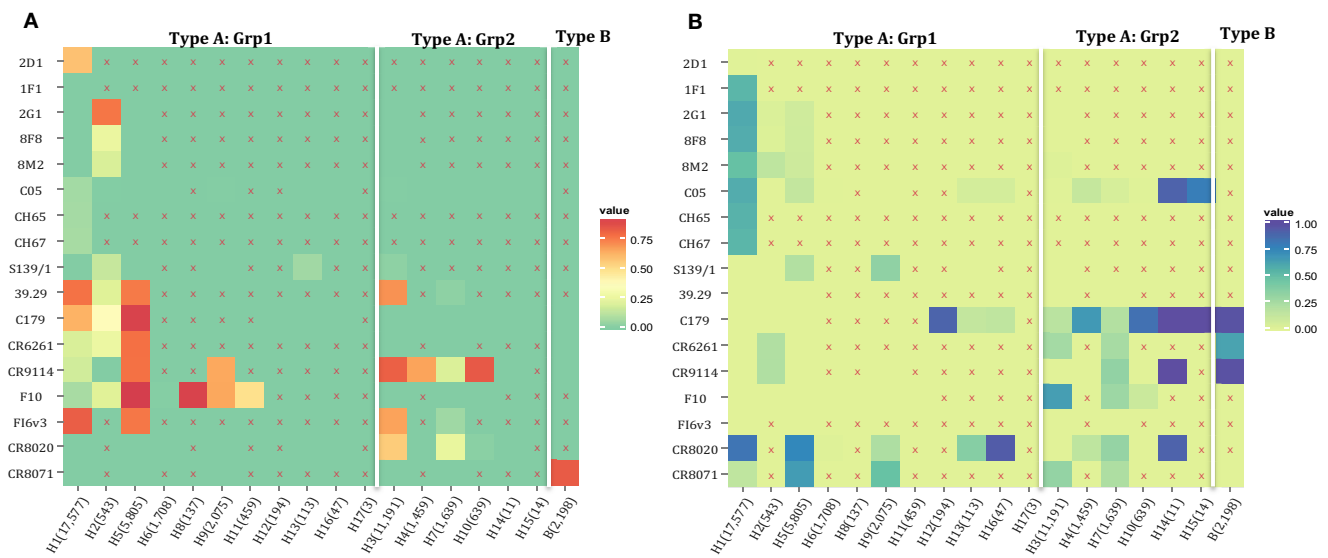


FIGURE 8 | The neutralized and escape motif coverage on each subtype within FLUKB of neutralizing antibodies. The heat map of **(A)** neutralized and **(B)** escape coverage are shown. Each cell on the grid represents coverage for a specific antibody (row name) on specific subtype (column name). The subtypes were sorted from Grp1 and Grp2 in influenza A virus, and influenza B virus. Different color schemes were used in order to differentiate neutralizing and escape coverage: from green to red/green to blue indicate rising neutralizing/escape coverage in HA sequence dataset. The boxes with symbol "X" indicate that no experimental validation data were available for this study.

Likewise, modification of glycosylation sites through sequence change may impact accessibility of antibodies to the neutralization site, creating discordance between sequence identity of binding site shown in BlockLogo and neutralization outcome between two strains of viruses sharing the same epitope sequence. The frequency of such occurrences will be important to determine. Neutralization assays of strains with discontinuous epitopes identical to validated B-cell epitopes will provide a proof of cross-neutralization. Since the experimental validation is time and money consuming, the introduction of extended B-cell epitope (see Supplementary Material) aims to help select representative sequences that differ in extended B-cell epitopes. For each proposed neutralizing or escape peptide (actual B-cell epitope), a small number of variants defined by changes in its environment

(extended B-cell epitopes) constitute the majority of strains with the proposed peptide.

On the other hand, before more experimental data generated to fill the existing "non-validated gap," it will be meaningful to bring out some reasonable estimation. The assumption and methods in this paper are based on complete identity to discontinuous motifs on B-cell epitope (additionally extended B-cell epitope). To check the validity of this assumption, the similarity between discontinuous motifs and discontinuous peptides could be used to estimate and predict neutralization and binding results in the future. For example, a discontinuous peptide with mutated residues of similar feature to the neutralized motif would be considered as "possible neutralized peptide" against specific nAbs. These estimations could also be validated in experimental assays, and then be used to further experimental design iteratively.

Table 5 | Distribution of the number of neutralizing antibodies that share identical neutralized discontinuous motif with sequences within the HA sequence dataset.

Number of nAbs	Coverage in 45,812 HA dataset (%)
0	17.38
1	12.45
2	11.68
3	13.39
4	31.38
5	1.59
6	9.89
7	2.25

For each sequence within the 45,812 HA dataset, the number of nAbs that share identical neutralized motif was counted. The number of nAbs in our panel for any given influenza strain can range from 0 to 7.

CONCLUSION

Over the past few years, our understanding of nAbs and their responses against influenza HA have expanded tremendously. Besides the well-known HA head region interactions, an increasing number of characterized nAbs bind and neutralize influenza virus by targeting the more conserved stem regions. Among these stem-targeting nAbs, some show broadly neutralizing ability across subtypes/lineages, even across two groups in influenza A virus strains. However, the related experimental data for majority of nAbs are quite limited.

In sum, we have established a library of validated motifs (extracted from HA sequences in neutralized and escape strains) for each nAb. For any newly emerging strain, the cross-neutralization prediction can be made rapidly for existing nAbs and validation experiments can be designed judiciously. This study provides a method for investigation of cross-reactivity of nAbs against influenza viruses, but is directly applicable to any viral pathogen that has structurally characterized nAbs and a collection of variant sequences of the target antigen. Examples of such pathogens include orthomyxoviruses (influenza); flaviviruses such as dengue or West Nile; arenaviruses such as

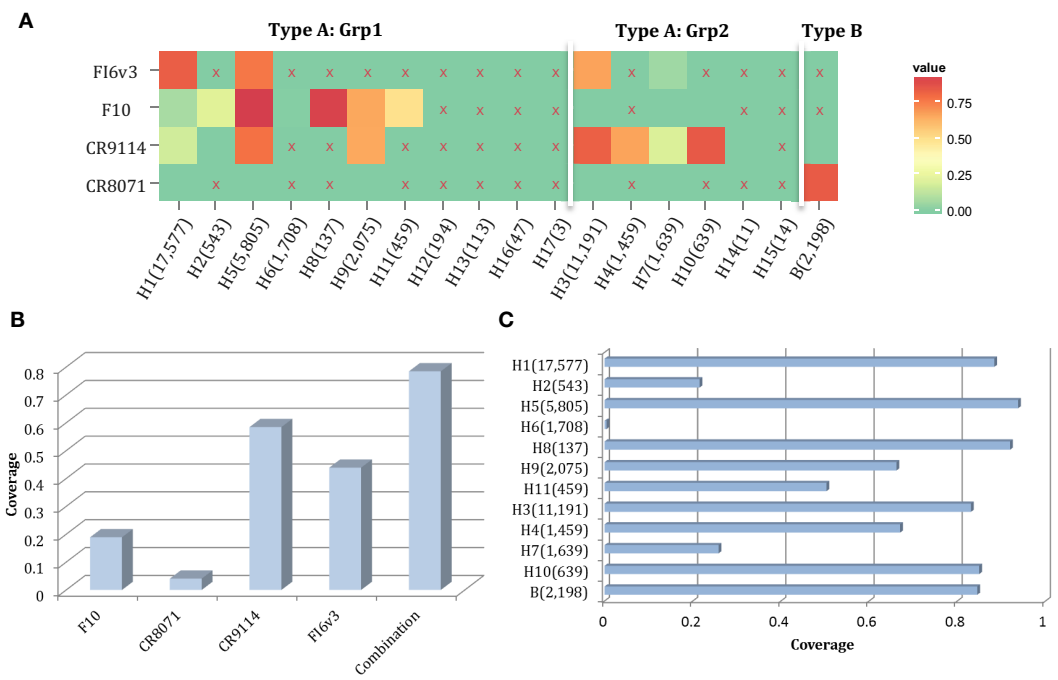


FIGURE 9 | A combination of four neutralizing subtype-diversified and potent neutralizing antibodies. **(A)** The heat map of the neutralized result for four nAbs, the color scheme is same as **Figure 8A**; **(B)** the neutralized coverage of four nAbs individually, and the combination of nAbs on HA dataset; and **(C)** the neutralizing result of combination of four nAbs by subtype. The subtypes were sorted from Grp1 and Grp2 in influenza A virus and influenza B virus. Only subtypes with neutralizing data are shown.

lymphocytic choriomeningitis virus and human immunodeficiency virus, among others. Insights from such bioinformatics analyses coupled with antibody antigenicity through crystallographic determinations will facilitate electronic neutralization profiling that can be tested empirically in subsequent laboratory neutralization assays.

ACKNOWLEDGMENTS

Jing Sun, Vladimir Brusic, and Ellis L. Reinherz acknowledge funding from NIH grant U01 AI 90043. Ulrich J. Kudahl was funded by Oticon Foundation, Otto Mønsted Foundation, Julie Dam's Stipend, Henry Shaws Stipend, and Reinholdt Jorcks Stipend. Christian Simon was funded by the Novo Scholarship Programme, Direktør Ib Henriksens Fond, and Augustinus Fonden.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00038/abstract

REFERENCES

1. Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, et al. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA* (2003) **289**(2):179–86. doi:10.1001/jama.289.2.179
2. Parvin JD, Moscona A, Pan WT, Leider JM, Palese P. Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J Virol* (1986) **59**(2):377–83.
3. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* (2009) **324**(5934):1557–61. doi:10.1126/science.1176062
4. Wiley DC, Skehel JJ. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu Rev Biochem* (1987) **56**:365–94. doi:10.1146/annurev.bi.56.070187.002053
5. Nakamura G, Chai N, Park S, Chiang N, Lin Z, Chiu H, et al. An in vivo human-plasmablast enrichment technique allows rapid identification of therapeutic influenza a antibodies. *Cell Host Microbe* (2013) **14**(1):93–103. doi:10.1016/j.chom.2013.06.004
6. Steel J, Lowen AC, Wang TT, Yondola M, Gao Q, Haye K, et al. Influenza virus vaccine based on the conserved hemagglutinin stalk domain. *MBio* (2010) **1**(1):e18–10. doi:10.1128/mBio.00018-10
7. Okuno Y, Isegawa Y, Sasao F, Ueda S. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. *J Virol* (1993) **67**(5):2552–8.
8. Yoshida R, Igarashi M, Ozaki H, Kishida N, Tomabechi D, Kida H, et al. Cross-protective potential of a novel monoclonal antibody directed against antigenic site B of the hemagglutinin of influenza A viruses. *PLoS Pathog* (2009) **5**(3):e1000350. doi:10.1371/journal.ppat.1000350
9. Ueda M, Maeda A, Nakagawa N, Kase T, Kubota R, Takakura H, et al. Application of subtype-specific monoclonal antibodies for rapid detection and identification of influenza A and B viruses. *J Clin Microbiol* (1998) **36**(2):340–4.
10. Smirnov YA, Lipatov AS, Gitelman AK, Claas EC, Osterhaus AD. Prevention and treatment of bronchopneumonia in mice caused by mouse-adapted variant of avian H5N2 influenza A virus using monoclonal antibody against conserved epitope in the HA stem region. *Arch Virol* (2000) **145**(8):1733–41. doi:10.1007/s007050070088
11. Sakabe S, Iwatsuki-Horimoto K, Horimoto T, Nidom CA, Le M, Takano R, et al. A cross-reactive neutralizing monoclonal antibody protects mice from H5N1 and pandemic (H1N1) 2009 virus infection. *Antiviral Res* (2010) **88**(3):249–55. doi:10.1016/j.antiviral.2010.09.007
12. Ekiert DC, Wilson IA. Broadly neutralizing antibodies against influenza virus and prospects for universal therapies. *Curr Opin Virol* (2012) **2**(2):134–41. doi:10.1016/j.coviro.2012.02.005
13. Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen LM, et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat Struct Mol Biol* (2009) **16**(3):265–73. doi:10.1038/nsmb.1566
14. Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, Jarrossay D, et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* (2011) **333**(6044):850–6. doi:10.1126/science.1205669
15. Yamashita M, Krystal M, Fitch WM, Palese P. Influenza B virus evolution: co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. *Virology* (1988) **163**(1):112–22. doi:10.1016/0042-6822(88)90238-3
16. Rota PA, Wallis TR, Harmon MW, Rota JS, Kendal AP, Nerome K. Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. *Virology* (1990) **175**(1):59–68. doi:10.1016/0042-6822(90)90186-U
17. Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R, Ekiert DC, et al. Highly conserved protective epitopes on influenza B viruses. *Science* (2012) **337**(6100):1343–8. doi:10.1126/science.1222908
18. Huang J, Honda W. CED: a conformational epitope database. *BMC Immunol* (2006) **7**:7. doi:10.1186/1471-2172-7-7
19. Squires RB, Noronha J, Hunt V, Garcia-Sastre A, Macken C, Baumgarth N, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respir Viruses* (2012) **6**(6):404–16. doi:10.1111/j.1750-2659.2011.00331.x
20. Tsibane T, Ekiert DC, Krause JC, Martinez O, Crowe JE Jr, Wilson IA, et al. Influenza human monoclonal antibody 1F1 interacts with three major antigenic sites and residues mediating human receptor specificity in H1N1 viruses. *PLoS Pathog* (2012) **8**(12):e1003067. doi:10.1371/journal.ppat.1003067
21. Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE Jr, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science* (2010) **328**(5976):357–60. doi:10.1126/science.1186430
22. Fleury D, Daniels RS, Skehel JJ, Knossow M, Bizebard T. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* (2000) **40**(4):572–8. doi:10.1002/1097-0134(20000901)40:4<572::AID-PROT30>3.3.CO;2-E
23. Ekiert DC, Kashyap AK, Steel J, Rubrum A, Bhabha G, Khayat R, et al. Cross-neutralization of influenza A viruses mediated by a single antibody loop. *Nature* (2012) **489**(7417):526–32. doi:10.1038/nature11414
24. Whittle JR, Zhang R, Khurana S, King LR, Manischewitz J, Golding H, et al. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proc Natl Acad Sci U S A* (2011) **108**(34):14216–21. doi:10.1073/pnas.1111497108
25. Schmidt AG, Xu H, Khan AR, O'Donnell T, Khurana S, King LR, et al. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc Natl Acad Sci U S A* (2013) **110**(1):264–9. doi:10.1073/pnas.1218256109
26. Bizebard T, Gigant B, Rigolet P, Rasmussen B, Diat O, Bosecke P, et al. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. *Nature* (1995) **376**(6535):92–4. doi:10.1038/376092a0
27. Fleury D, Barrere B, Bizebard T, Daniels RS, Skehel JJ, Knossow M. A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. *Nat Struct Biol* (1999) **6**(6):530–4. doi:10.1038/9299
28. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, et al. An antibody that prevents the hemagglutinin low pH fusogenic transition. *Virology* (2002) **294**(1):70–4. doi:10.1006/viro.2001.1320
29. Lee PS, Yoshida R, Ekiert DC, Sakai N, Suzuki Y, Takada A, et al. Heterosubtypic antibody recognition of the influenza virus hemagglutinin receptor binding site enhanced by avidity. *Proc Natl Acad Sci U S A* (2012) **109**(42):17040–5. doi:10.1073/pnas.1212371109
30. Dreyfus C, Ekiert DC, Wilson IA. Structure of a classical broadly neutralizing stem antibody in complex with a pandemic H2 influenza virus hemagglutinin. *J Virol* (2013) **87**(12):7149–54. doi:10.1128/JVI.02975-12
31. Throsby M, van den Brink E, Jongeneelen M, Poon LL, Alard P, Cornelissen L, et al. Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PLoS One* (2008) **3**(12):e3942. doi:10.1371/journal.pone.0003942
32. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, et al. Antibody recognition of a highly conserved influenza virus epitope. *Science* (2009) **324**(5924):246–51. doi:10.1126/science.1171491

33. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, et al. A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* (2011) **333**(6044):843–50. doi:10.1126/science.1204839
34. Sun X, Shi Y, Lu X, He J, Gao F, Yan J, et al. Bat-derived influenza hemagglutinin H17 does not bind canonical avian or human receptors and most likely uses a unique entry mechanism. *Cell Rep* (2013) **3**(3):769–78. doi:10.1016/j.celrep.2013.01.025
35. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* (2007) **23**(21):2947–8. doi:10.1093/bioinformatics/btm404
36. Page RD. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* (1996) **12**(4):357–8.
37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res* (2000) **28**(1):235–42. doi:10.1093/nar/28.1.235
38. Katz J, Hancock K, Veguilla V, Zhong W, Lu X, Sun H, et al. Serum cross-reactive antibody response to a novel influenza A (H1N1) virus after vaccination with seasonal influenza vaccine. *MMWR Morb Mortal Wkly Rep* (2009) **58**(19):521–4.
39. Stavitsky AB. Micromethods for the study of proteins and antibodies II. Specific applications of hemagglutination and hemagglutination-inhibition reactions with tannic acid and protein-treated red blood cells. *J Immunol* (1954) **72**(5):368–75.
40. Yu X, Tsibane T, McGraw PA, House FS, Keefer CJ, Hicar MD, et al. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* (2008) **455**(7212):532–6. doi:10.1038/nature07231
41. Krause JC, Tumpey TM, Huffman CJ, McGraw PA, Pearce MB, Tsibane T, et al. Naturally occurring human monoclonal antibodies neutralize both 1918 and 2009 pandemic influenza A (H1N1) viruses. *J Virol* (2010) **84**(6):3127–30. doi:10.1128/JVI.02184-09
42. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* (2008) **9**(4):286–98. doi:10.1093/bib/bbn013
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* (1997) **25**(17):3389–402. doi:10.1093/nar/25.17.3389
44. Hubbard SJ, Thornton JM. *Naccess. Computer Program.* (Vol. 2). London: Department of Biochemistry and Molecular Biology, University College London (1993).
45. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. *Proc Natl Acad Sci U S A* (2003) **100**(6):3215–20. doi:10.1073/pnas.0535768100
46. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* (2007) **372**(3):774–97. doi:10.1016/j.jmb.2007.05.022
47. Wilson IA, Skehel JJ, Wiley DC. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. *Nature* (1981) **289**(5796):366–73. doi:10.1038/289366a0
48. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* (2004) **14**(6):1188–90. doi:10.1101/gr.849004
49. Olsen LR, Kudahl UJ, Simon C, Sun J, Schönbach C, Reinherz EL, et al. BlockLogo: Visualization of peptide and sequence motif conservation. *J Immunol Methods* (2013) **400**:37–44. doi:10.1016/j.jim.2013.08.014

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 September 2013; accepted: 22 January 2014; published online: 07 February 2014.

*Citation: Sun J, Kudahl UJ, Simon C, Cao Z, Reinherz EL and Brusic V (2014) Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin – implications for cross-reactivity of neutralizing antibodies. Front. Immunol. **5**:38. doi: 10.3389/fimmu.2014.00038*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Sun, Kudahl, Simon, Cao, Reinherz and Brusic. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 December 2013
doi: 10.3389/fimmu.2013.00427

Pre-clustering of the B cell antigen receptor demonstrated by mathematically extended electron microscopy

Gina J. Fiala*[1,2,3†], *Daniel Kaschek*[4†], *Britta Blumenthal*[1,5], *Michael Reth*[1,3,6], *Jens Timmer*[3,4] and *Wolfgang W. A. Schamel*[1,3,5]

[1] *Faculty of Biology, Department of Molecular Immunology, Albert Ludwigs University Freiburg, Freiburg, Germany*
[2] *Spemann Graduate School of Biology and Medicine (SGBM), Albert Ludwigs University Freiburg, Freiburg, Germany*
[3] *Centre for Biological Signalling Studies BIOSS, Albert Ludwigs University Freiburg, Freiburg, Germany*
[4] *Institute of Physics, Albert Ludwigs University Freiburg, Freiburg, Germany*
[5] *Medical Faculty, Centre for Chronic Immunodeficiency CCI, University Clinics Freiburg, Albert Ludwigs University Freiburg, Freiburg, Germany*
[6] *Max Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Shiv Pillai, Harvard Medical School, USA
To-Ha Thai, Beth Deaconess Israel Medical Center, USA

****Correspondence:***
Wolfgang W. A. Schamel, Faculty of Biology, Department of Molecular Immunology, Albert Ludwigs University Freiburg, Schänzle Straße 18, Freiburg 79104, Germany
e-mail: wolfgang.schamel@biologie.uni-freiburg.de

[†] *Gina J. Fiala and Daniel Kaschek have contributed equally to this work.*

The B cell antigen receptor (BCR) plays a crucial role in adaptive immunity, since antigen-induced signaling by the BCR leads to the activation of the B cell and production of antibodies during an immune response. However, the spatial nano-scale organization of the BCR on the cell surface prior to antigen encounter is still controversial. Here, we fixed murine B cells, stained the BCRs on the cell surface with immuno-gold and visualized the distribution of the gold particles by transmission electron microscopy. Approximately 30% of the gold particles were clustered. However the low staining efficiency of 15% precluded a quantitative conclusion concerning the oligomerization state of the BCRs. To overcome this limitation, we used Monte-Carlo simulations to include or to exclude possible distributions of the BCRs. Our combined experimental-modeling approach assuming the lowest number of different BCR sizes to explain the observed gold distribution suggests that 40% of the surface IgD-BCR was present in dimers and 60% formed large laminar clusters of about 18 receptors. In contrast, a transmembrane mutant of the mIgD molecule only formed IgD-BCR dimers. Our approach complements high resolution fluorescence imaging and clearly demonstrates the existence of pre-formed BCR clusters on resting B cells, questioning the classical cross-linking model of BCR activation.

Keywords: BCR, oligomerization, electron microscopy, immuno-gold-labeling, Monte Carlo simulation, maximum-likelihood method

1. INTRODUCTION

Cells communicate with each other and with their surroundings through transmembrane receptors that are embedded in the plasma membrane. Thus, it is of high interest to understand how these receptors and other cell surface proteins, such as adhesion molecules or channels, are organized on the membrane. Initially it was thought that proteins and lipids freely diffuse in membranes and that they are randomly distributed (1). With the concept of lipid rafts it was noted that specialized microdomains on the cell surface exist, where some proteins are concentrated and others are excluded (2). Although the raft concept had to be modified, since in biological membranes they are smaller and more transient than in artificial model membranes (3), it is clear that proteins are not randomly distributed on the cell surface. One example of the immune system is the T cell antigen receptor that can form pre-clustered oligomers, called nanoclusters, on T cells (4–7). Nanoclusters form before and independently of any ligand encounter. Interestingly, T cells can control the degree of TCR nanoclustering, in order to regulate their avidity toward multivalent ligands and thus their sensitivity (8–10). This indicates that studying the nano-scale distribution of a receptor contributes to the understanding of the function of the receptor. Less well understood is a potential pre-clustering of the B cell antigen receptor (BCR). The BCR is expressed on B cells and controls the development of these cells and their activation upon contact with the BCR's ligand, called antigen. The BCR is composed of the membrane-bound immunoglobulin (mIg) molecule and a heterodimer of the Igα (mb-1) and Igβ (B29) proteins (11). The mIg molecule binds to the antigen and exists in different isotypes, of which the mIgD form is the most abundant one on resting mature B cells (12, 13). The Igα/Igβ dimer contains phosphorylatable tyrosines in the cytoplasmic tails (14) and transmits the signal of antigen-binding to the cytoplasmic signaling machinery. The first evidence for BCR pre-clustering, i.e., the existence of BCR oligomers, was obtained by Blue Native gel electrophoresis (15–17). Upon extraction of the IgD- and IgM-BCRs from the cell membrane of resting B cells using low concentrations of detergent, the BCRs were found in oligomers. Importantly, a mutant mIgD molecule, in which the transmembrane region was mutated (called mIgD-hSbap), only formed dimers (16). Thus, in the BCR as well as in the TCR (10), the transmembrane region of the ligand-binding subunits is involved in the pre-clustering. Later, three approaches were used to investigate whether the BCR forms oligomers in living cells. Firstly, a FRET approach was used, and oligomers were

not detected (18). Secondly, a bifluorescence complementation approach was used, and BCR oligomers were detected (19). The differences in these two procedures and possible functional implications of pre-clustered BCR oligomers were recently discussed (20). Thirdly, the superresolution microscopy method direct stochastic optical reconstruction microscopy (dSTORM) was used to show that BCRs were organized as pre-clusters on the surface of resting primary B cells (21). Here, we used a validated technique, that we had previously used to study TCR pre-clustering (6, 8), in order to answer the question of whether BCR oligomers exist and if yes what their sizes are. To this end, we used fixed B cells and labeled the BCRs with specific antibodies that were bound to gold particles (immuno-gold-staining), prepared cell surface replicas and analyzed the nano-scale distribution of the gold particles by transmission electron microscopy (TEM). This approach allowed visualization of BCRs on cell surface areas that do not adhere to any experimental support, giving the opportunity to analyze untouched (and non-modified) receptors. A general challenge of immuno-gold-labeling is its low staining efficiency, which is compensated by mathematical modeling and statistical methods, allowing solid conclusions to be drawn from the experimental data.

2. MATERIALS AND METHODS

2.1. EXPERIMENTAL PROCEDURES

2.1.1. Cell culture and cell fixation

The murine B cell lines J558L (not expressing any BCR) and J558Lδm/mb-1flN (expressing an IgD-BCR) were previously described (16). We also used a J558L line that expressed a mutant IgD-BCR, in which the transmembrane region of the mIgD molecule was mutated (mIgD-hSbap) (16). Cells were cultured in RPMI 1640 complete medium supplemented with 10% fetal calf serum, 2 mM L-glutamine, 100 U/ml penicillin/streptomycin, 10 mM HEPES, and 50 mM 2-mercaptoethanol and grown at 37°C in a humidified atmosphere with 5% CO2. Cells were fixed with freshly prepared, ice-cold 4% paraformaldehyde in PBS for 20 min at 4°C at a cell density of 10×10^6 cells/ml. After fixing, cells were washed twice with cold PBS.

2.1.2. Immuno-gold-staining and analysis of gold-reagent

Unstimulated PFA-fixed cells were stained with the primary anti-idiotypic antibody Ac146 (22) at saturating concentration of 20 mg/ml in PBS with 1% BSA for 1 h on ice. This antibody binds to the variable regions of the BCR used in this study. Staining with a secondary anti-mouse IgG antibody conjugated to 10 nm gold (Aurion) was performed for 1 h on ice. Prior to cell staining, the aggregation state of the gold-reagent was tested by adsorbing diluted suspensions of the gold-reagent onto collodion/carbon-coated EM grids, which were analyzed in transmission electron microscopy (TEM, **Figure 2**).

2.1.3. Surface replica preparation

Labeled cells were adsorbed to L-poly-lysine-treated micas, followed by a second fixation with 0.1% glutaraldehyde on ice for 30 min. Micas containing stained cells were covered with an untreated piece of mica and fast-frozen in a Reichert-Jung (now Leica) KF-80 plunge freezing unit using the secondary cryogen liquid ethane. Metal replicas were prepared in a freeze fracture unit (BAF 060; BAL-TEC) where the cell-containing mica slide was freeze-etched at −150°C for 12 min to sublime surface ice. Frozen cells were then shadowed with 2 nm of evaporated platinum at an angle of 45°C and strengthened by a uniformly thick 20 nm electron-translucent carbon layer evaporated perpendicular to the mica surface plane. The metal replica of the surface was released from the mica by floating it on commercial bleach where it remained over night for digestion of the organic material. The floating replica was washed three times in distilled water to remove attached organic material and chemicals and then picked up on uncoated copper EM grids. For a detailed protocol see Ref. (23).

2.1.4. Analysis of metal replicas by TEM

Replicas mounted on EM grids were examined in a transmission electron microscope (1200-EX II; JOEL) operating at 100 kV. Gold particle numbers and the gold cluster size distribution were counted and analyzed for at least 3 cells per sample at an augmentation of 25000. Gold particles were considered to be part of the same cluster when they were adjacent or less distant than 10 nm (the diameter of a single gold particle), taking into account that the diameter of the BCR is around 10 nm (24). Pictures were taken at augmentations of 5000 (cell overview), 120000, and 300000 (gold-labeling).

2.1.5. Quantification of the number of BCRs per cell

The number of BCRs per cell was determined based on a saturation binding assay. About 1×10^6 J558Lδm/mb-1flN cells were stained for 30 min at 4°C with increasing concentrations of an FITC-coupled anti-IgD antibody (BD Pharmingen, clone 11-26c.2a). Following 5 extensive washing steps, the fluorescence signal was measured in duplicates using a SpectraMax 190 Absorbance Microplate Reader. In order to convert the fluorescence signal intensity into the number of antibodies, a calibration of the antibody was performed by fitting the model $y = mx + b$ to the standard concentrations x and fluorescence signals y. Here m denotes the slope and b denotes the intercept of the calibration curve. In order to infer the number of receptors on the cells, antibody was spotted in different concentrations. The saturation model $y = y_0 + \frac{ax}{b+x}$ was fitted to the sample data. Here, x denotes the antibody concentration and y denotes the fluorescence signal after washing. The parameter y_0 was measured explicitly. The remaining parameters, i.e., the maximal fluorescence signal gain a and the saturation constant b were estimated from the data. From the maximum signal gain a, the corresponding concentration of bound antibody was computed from the calibration curve, i.e., $\Delta x = \frac{a}{m}$. Finally, the number of receptors per cell was computed by the formula $n = \frac{\Delta x \cdot N_A \cdot V}{k \cdot N_{\text{cells}} \cdot M}$, where $N_A = 6.022 \times 10^{23}$ 1/mol, $V = 50\,\mu$l, $k = 2$, $N_{\text{cells}} = 5 \times 10^5$, and $M = 146.389 \times 10^3$ g/mol denote Avogadros constant, the volume per well, the number antibody binding sites per receptor, the number of cells, and the molecular weight of the antibody. From the parameters obtained by the calibration and saturation curve, we get $n = 122400 \pm 7500$. The uncertainty of the number of receptors per cell is dominated by the uncertainty of a, the maximum signal gain. The uncertainty of a is propagated to the error of n by Gaussian error propagation.

2.2. MONTE-CARLO SIMULATION OF OBSERVED GOLD CLUSTER SIZE DISTRIBUTION

The immuno-gold-staining and counting process was simulated by a Monte-Carlo approach. It is assumed that the observed gold cluster size distribution is a superposition of distributions generated by single size oligomers. Each oligomer size produces a characteristic distribution of observed gold cluster sizes that depends on the staining efficiency. The characteristic distribution ranges from exclusively monomeric observation to exclusively single size oligomeric observation for staining probabilities zero and one, respectively. The distributions in between zero and one depend on the oligomer geometry, which is reflected by the number of next neighbors of an average receptor. This number is at least 2, i.e., for linear arrangement of the receptors. For other cases, like dense circle packing resulting in a triangular geometry it is 6 and for a quadratic grid it is 8. Simulations have been performed for linear arrangements and quadratic grids which reflect different extremes. An additional factor for the observed size distribution is the gold-reagent itself, which is potentially pre-clustered. Further, the staining efficiency, i.e., the number of receptors that are stained; the geometry, i.e., the receptor positions within the oligomers being stained; and potential unspecific stainings, i.e., presence of gold particles that are not bound to any BCR, have to be considered. For given oligomer size, staining efficiency, geometry, and gold distribution, the observed gold cluster size distribution is obtained by repeated random number generation for the number of stained receptors, their positions and the number of gold particles per staining spot. For each set of random numbers, the resulting representation of the gold particle pattern is evaluated by the simulation program and the number of counted monomers, dimers, etc., is collected. This procedure was performed 10^5 times for 10 staining probabilities between 2 and 40%, underlying BCR oligomer sizes from 1 to 40 and three geometries, i.e., linear, triangular, and quadratic. In addition, the simulation approach has been adapted to explain the observation of the gold-reagent control experiment.

2.3. STATISTICAL INFERENCE

The result of each gold-staining experiment is a distribution of gold cluster sizes determined from gold particle counting in the microscope. These experimental data are compared to the simulated data and by means of statistical methods it is decided whether simulation and experiment are in accordance. We tested four major hypotheses: receptors are organized as:

1. BCR oligomers of a unique fixed size s,
2. BCR monomers and oligomers of a unique fixed size s,
3. BCR dimers and oligomers of a unique fixed size s,
4. BCR monomers, dimers, and oligomers of a unique fixed size s.

For each hypothesis, a likelihood function is derived based on the assumption of Poisson statistics for the counted gold oligomers. The corresponding log-likelihood reads

$$l(\theta) = \sum_i m_i(\theta) - N_i \log(m_i(\theta)) + \sum_{k=1}^{N_i} \log\ k, \qquad (1)$$

where m_i is the number of expected BCR oligomers of size i predicted by the simulation and N_i is the number of gold oligomers counted in the experiment. Minimization is performed with respect to the parameter vector θ which is defined differently for each hypothesis. This is explicitly

$$m_i^{(1)}(\theta) = \theta \cdot n_{i,sim}(s), \qquad (2)$$

$$m_i^{(2)}(\theta) = \theta_1 \cdot n_{i,sim}(1) + \theta_2 \cdot n_{i,sim}(s), \qquad (3)$$

$$m_i^{(3)}(\theta) = \theta_1 \cdot n_{i,sim}(2) + \theta_2 \cdot n_{i,sim}(s), \qquad (4)$$

$$m_i^{(4)}(\theta) = \theta_1 \cdot n_{i,sim}(1) + \theta_2 \cdot n_{i,sim}(2) + \theta_3 \cdot n_{i,sim}(s), \qquad (5)$$

where $n_{i,sim}(s)$ denotes the simulated observed gold cluster size distribution given the underlying BCR oligomer size. I.e., the parameter vector θ reflects the composition of BCR oligomers on the surface. For each simulated data set, i.e., for each triplet (staining efficiency p, underlying cluster size s, geometry g), the maximization of the log-likelihood results in an estimate for the parameter vector θ, denoted by $\hat{\theta}$, that explains the experimental data best. In addition, to test if the log-likelihood value $l(\hat{\theta})$ is in accordance with the data, parametric bootstrapping is employed. In parametric bootstrapping, the model prediction $m_i(\hat{\theta})$ is used to generate *de novo* observation data. In our case, 10^3 random samples have been drawn from a Poisson distribution with mean $m_i(\hat{\theta})$ and have been treated like gold particle observations. For each sample, the log-likelihood is maximized again and the values are collected in a histogram approximating the asymptotic log-likelihood distribution. The original value $l(\hat{\theta})$ is compared to different statistics of the sampled distribution, among these p-value and weighted distance to the mean $\Delta = \frac{l(\hat{\theta}) - \langle l(\theta) \rangle}{\sigma_{l(\theta)}}$. The hypotheses are rejected based on these values for $p < 0.01$ and $\Delta > 3$, respectively.

3. RESULTS

3.1. THE WORKFLOW

Here we derived the size distribution of the IgD-BCRs on the cell surface of J558L transfectants from the measured distribution of gold particles after staining the BCRs with immuno-gold. "Size distribution of the BCRs" is defined as the percentage of BCRs in a given cluster size, such as 5% of the BCRs are in BCR monomers, 45% are in BCR dimers, and 50% are in BCR trimers. Our workflow consists of two phases. The first is carried out in the wet lab and comprises immuno-gold-labeling of the BCRs on untreated and PFA-fixed cells, followed by cell surface replica preparation and the quantification of the gold particle cluster size distribution on the replicated cell surface area by TEM (**Figure 1**, steps 1 and 2). As the staining efficiency reached by immuno-gold-labeling is low, the obtained immuno-gold data cannot directly be converted into the distribution of the BCRs. Thus, in the second phase mathematical modeling is used to derive the BCR distribution from the gold particle data. To do so, we assume a large number of different BCR distributions, such as only one defined size (only monomers, or only dimers, or only trimers, etc.) or a combination of sizes (for example, dimers and trimers). The observed gold particle size distributions are then stochastically simulated using a Monte-Carlo simulation of the staining process, and the likelihood that the gold size distribution represents a given BCR size distribution is

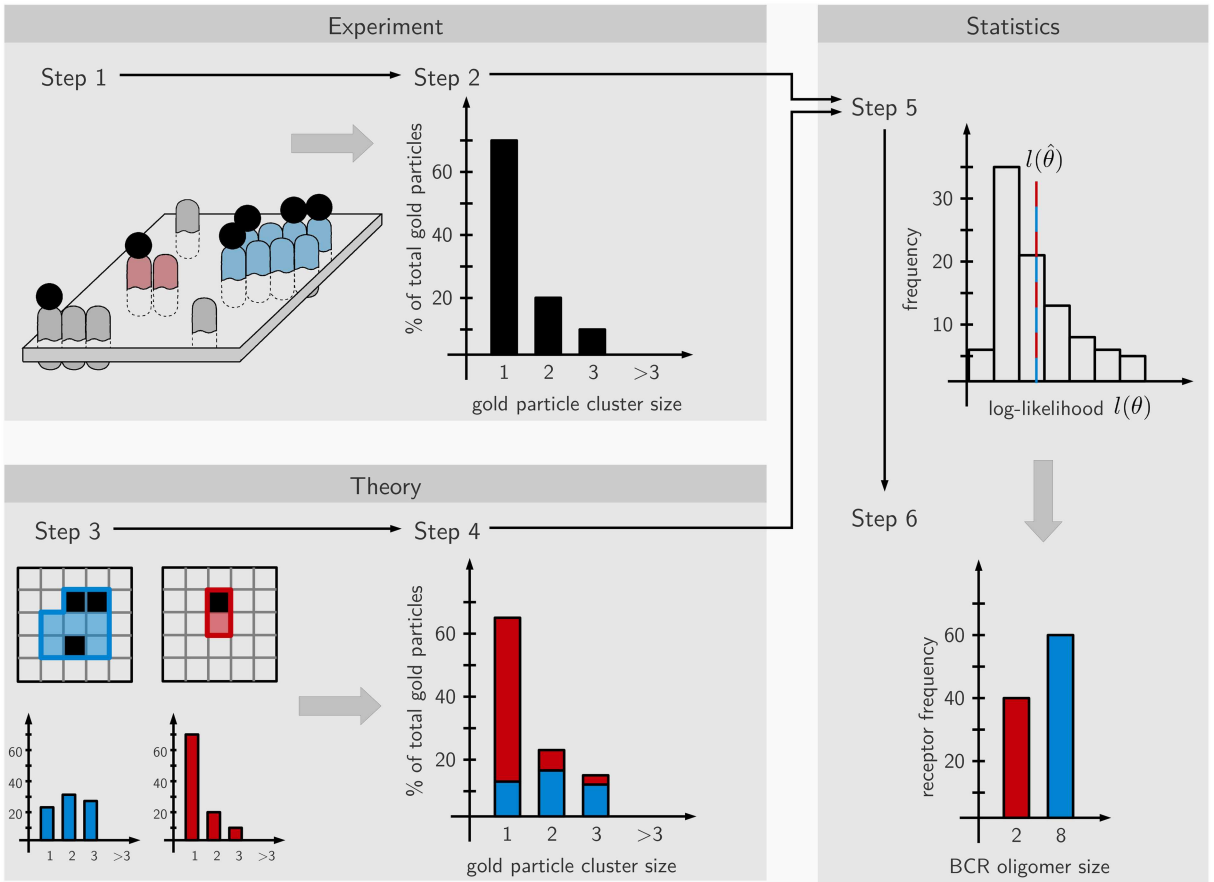


FIGURE 1 | Scheme of the workflow. The approach is divided into six steps partitioned into experimental (steps 1 and 2), theoretical (steps 3 and 4), and statistical parts (steps 5 and 6). In step 1, the IgD-BCRs on the surface of resting and fixed B cells are stained with immuno-gold and the gold particles visualized by TEM. In step 2, the gold particles are counted, generating an observed gold cluster size distribution. In step 3, the staining process is simulated by Monte-Carlo procedures based on hypothetical BCR nanocluster size distributions, generating characteristic gold particle cluster distributions for the observed gold-labeling. Gold particles are shown in black and BCRs in blue or red (each BCR occupies one square, thus showing a blue 8 mer and a red dimer). In step 4, superposition parameters are estimated based on the observed gold particle distributions by maximum-likelihood estimation. In step 5 based on the estimated parameters, *de novo*, *in silico* cluster size distributions are repeatedly generated and the corresponding log-likelihood is re-optimized, providing a histogram of log-likelihood values. Step 6; if in accordance with the data, the theory makes a valid prediction for the underlying BCR nanocluster distribution and the relative amounts of the nanocluster sizes is calculated.

calculated. Each single BCR oligomer size produces a characteristic distribution of observed gold cluster sizes that depends on (1) the pre-clustering of the gold-reagent, i.e., presence of monomeric, dimeric, or trimeric gold particles in the staining reagent, (2) the immuno-gold-staining efficiency, i.e., the percentage of BCRs that are labeled by gold particles, (3) unspecific binding of the gold-reagent to the B cell, independent of any BCR, (4) the oligomer geometry which is reflected by the number of next neighbors of an average receptor within the BCR oligomer, this number is at least 2, i.e., for linear chains of receptors, or up to 8 for a quadratic grid (**Figure 3A**). The Monte-Carlo simulation provides information on the staining patterns and thus, the visible gold cluster size distribution (step 3). To match the data, several such gold cluster distributions are superposed with different strengths. These strengths, called the superposition parameters, are estimated from the data and give rise to a hypothesized underlying BCR size distribution (step 4). Based on the estimated superposition parameters the gold particle observation data is repeatedly generated and the parameters are re-estimated. This process results in a distribution of log-likelihood values which, in case of good agreement between model and data, contains the original log-likelihood value $l(\hat{\theta})$ (step 5). For the log-likelihood value, the threshold corresponding to a p-value of 0.01 can be computed which when being exceeded allows to reject the model. Otherwise, for a model log-likelihood value in the interior of the distribution, the model makes a valid prediction for the underlying BCR oligomer size distribution (step 6). This means that the observed gold cluster distribution can be fully explained by the predicted BCR oligomer distribution.

3.2. ESTIMATION OF THE DEGREE OF PRE-CLUSTERING OF THE GOLD-REAGENT

To address a possible pre-clustering of the gold-reagent used for the BCR immuno-gold-labeling, the gold-reagent alone was adsorbed

onto collodion/carbon-coated EM grids, i.e., without receptor binding, and analyzed by TEM (**Figure 2A**). The few observed gold particle dimers and trimers could be due to either pre-clustering of the gold particles in the staining reagent or random collocation of monomeric gold particles on the grid. The observed size distribution data of the gold-reagent alone (**Figure 2B**) was subjected to our workflow described above and interpreted as the staining of one huge quadratic oligomer. The size of this quadratic oligomer is the number of image pixels divided by the number of pixels per gold dot. The staining efficiency was first assessed by the number of gold particles divided by the oligomer size. Subsequently, the observed oligomer size distribution has been simulated by our Monte-Carlo approach showing that random collocation is not sufficient to explain the number of observed gold particle dimers but that the gold-reagent is indeed already pre-clustered. From this analysis, we find that 92.8% of all gold particles are monomeric, 6.5% are pre-clustered dimers, and 0.7% are pre-clustered trimers. These numbers have been taken into account for all following simulations by introducing parameters representing the inherent fraction of gold dimers and trimers. A complementary perspective on **Figure 2A** is its interpretation as staining of a surface with exclusively receptor monomers. This perspective enables insights into the specificity of the approach. Testing the hypothesis "100% monomers," differences in the staining efficiency should not lead to differences in the log-likelihood because only the total number of counts but not the observed size distribution changes. In contrast, when testing alternative hypotheses such as "100% dimers" or "100% trimers," the method should allow rejecting those hypotheses. Indeed, this was the case (**Figure 2C**). The plot shows the weighted distance to the mean and the p-value based on the log-likelihood. These values were computed for different hypothetical staining efficiencies, represented by different colors, and dominant oligomer sizes, represented by the x-axis. Already for hypothetical staining efficiencies larger than 3% all tested hypotheses other than "100% monomers" can be rejected with $p \leq 0.001$. Conversely, the monomer assumption is not rejected for any staining efficiency, in accordance with the expectation. This proofs that our approach is highly sensitive as a staining efficiency of 3% is already sufficient to reject wrong hypotheses.

3.3. CALCULATION OF THE IMMUNO-GOLD-STAINING EFFICIENCY FOR THE IgD-BCR

Next, we calculated the staining efficiency of the BCR immuno-gold-labeling process, using the murine B cell line J558Lm/mb-1flN (16). These cells express IgD-BCRs on their surface, as seen by a flow cytometric analysis using the same monoclonal anti-BCR antibody (Ac146) that was also used for the immuno-gold-labeling (**Figure 3B**). The total number of IgD-BCRs per cell was experimentally determined to be 122400 ± 7500 BCRs (described in the Methods section). The total cell surface area of a J558L cell is $782 \pm 25\,\mu m^2$ (25) resulting in an expected mean density of 156 BCRs/μm^2. We analyzed three J558Lδm/mb-1flN cells by immuno-gold-staining and TEM. The three areas analyzed by TEM were 101, 183, and 152 μm^2 and the expected BCR number in those areas was calculated to be 15600, 28080, and 23400 BCRs, respectively. The achieved BCR immuno-gold-labeling efficiency was calculated based on the number of gold particles observed in the areas (2192, 3577, and 3876 gold particles, respectively) assuming one gold particle to represent one BCR. Thus, the BCR labeling efficiencies were 14.1, 12.7, and 16.6%, respectively. Thus, the BCR staining efficiency in our experiment was approximately 15%.

3.4. THE ANTI-BCR IMMUNO-GOLD-STAINING IS SPECIFIC FOR THE BCR

To prove that the anti-BCR immuno-gold-labeling protocol only stained BCRs, we compared J558L cells that lack BCR expression

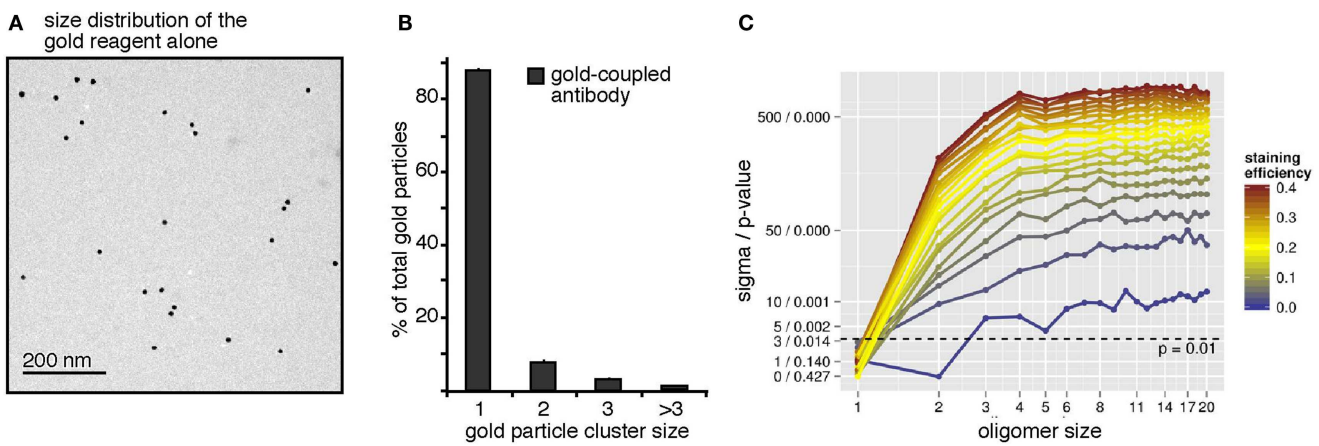


FIGURE 2 | The antibody-coupled gold-reagent is mainly monomeric. **(A)** The gold-reagent alone was adsorbed onto collodion/carbon-coated EM grids and analyzed by TEM. **(B)** Analysis of the clustering of 2041 gold particles alone showed 88% monomeric gold particles, 8% of gold particles as close as dimers, 3% were counted as trimers, and 1% as clusters of larger sizes. **(C)** The gold-reagent cluster size distribution was re-evaluated by our workflow, interpreting the cluster counts as receptor staining. Pre-clustering of the gold-reagent was taken into account. For different assumed receptor oligomer sizes (x-axis) and staining efficiencies (color scale), significance level and p-value were computed (y-axis). Models are rejected above the dashed line corresponding to a p-value of 0.01. The analysis confirms the monomer hypothesis and rejects other hypotheses with $p < 0.01$ (above the dashed line) for staining efficiencies larger 3%.

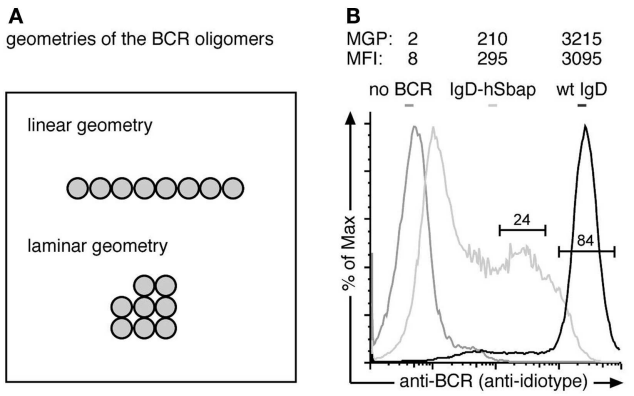


FIGURE 3 | The anti-BCR immuno-gold-labeling is specific. **(A)** The Monte-Carlo simulations were performed for two possible BCR oligomer geometries: a linear and a laminar geometry. **(B)** J558Lδm/mb-1flN cells expressing the wt IgD-BCR (black line), J558L cells without any BCR (dark gray line) or IgD-hSbap BCR-expressing J558L (light gray line) were labeled with the anti-BCR antibody Ac146 and subsequently stained with FITC-labeled anti-mouse IgG or gold-coupled anti-mouse IgG antibodies. Fluorescently labeled cells were analyzed by flow cytometry, and mean fluorescent intensities (MFI) of the gated populations are given. Gold labeled cells were subjected to replica preparation and analyzed by TEM. The mean gold particle (MGP) number detected on the cell replicas confirms specific labeling of the BCR on wt and IgD-hSbap expressing J558L cells, while immuno-gold-labeling of J558L not expressing BCR does not result in significant gold particle detection on surface replicas.

and the J558Lδm/mb-1flN cells (**Figure 3A**). Flow cytometry analysis confirmed no BCR expression on J558L cells. Immuno-gold-labeling of the J558L cells resulted in 2 gold particles per observed area, while J558Lδm/mb-1flN cells were stained with 3215 gold particles on average (**Figure 3A**). Thus (nearly) each gold particle is representing a BCR.

3.5. DETERMINATION OF THE SIZE DISTRIBUTION OF WT IgD-BCR OLIGOMERS

For our simulations to determine the size distribution of the wt IgD-BCR, we have used the immuno-gold data from J558Lδm/mb-1flN cells (**Figure 4A**) and the corresponding size distribution of the gold particles (**Figure 4B**). The quantification of gold clusters by TEM revealed 66% gold monomers, 21% gold in dimers, 7.5% gold in trimers, and a fraction of 5.5% gold particles were part of oligomers of four or more gold particles. To account for different possible geometries of each individual BCR cluster, simulations have been performed for linear and laminar BCR arrangements (**Figure 3B**). Firstly, we have assumed a linear arrangement of the BCRs (**Figure 4C**). Assuming the presence of one defined BCR oligomer size, labeled by x, we have performed simulations for BCR monomers (1), BCR dimers (2), BCR trimers (3), etc. (**Figure 4C**, top, left panel). The different staining probabilities are denoted by the color coding being centered around the experimentally measured value of 15%. All curves outside a range of $\pm$5% were shaded. For the single size assumption none of the assumed BCR cluster sizes is in agreement with the data (below the dotted line). The discrepancy between data and model does not change any more as soon as the assumed BCR oligomer size is larger than 6 because the observed size distribution produced by them is almost identical. In particular, this means that linear BCR oligomers larger than 6 or even a mixture of large BCR oligomers is not in accordance with the experimental data at the given staining efficiency. Also a mixture of small and large BCR oligomer sizes, e.g., $1 + x$ (BCR monomers and one BCR cluster size of x, upper, right panel), $2 + x$ (BCR dimers and one BCR cluster size of x, lower, left panel), or $1 + 2 + x$ (BCR monomers, dimers, and one BCR cluster size of x, lower, right panel), is unable to explain the observed data by linearly arranged receptor oligomers, thus, leading to a full rejection of this geometry hypothesis. Secondly, we have assumed a laminar arrangement of the BCRs (**Figure 4D**). Following the plot for the single size assumption (**Figure 4D**, panel x), the best agreement is obtained for small BCR oligomers such as BCR dimers and trimers. This suggests that small BCR oligomers contribute significantly to the overall observed gold cluster distribution. Evaluation of related model hypotheses, $1 + x$, $2 + x$, and $1 + 2 + x$ shows that monomers plus an additional BCR oligomer size is not sufficient to explain the observed size distribution. However, accordance can be achieved for BCR dimers plus a large BCR oligomer around a size of 18 (lower, left panel). Additional BCR monomers do not further reduce the discrepancy between model and data (lower, right panel). This is also expressed in **Figure 4F**. BCR dimers dominate the small observed gold particle oligomers up to a size of 3, where BCR dimers are occasionally observed as trimers due to the pre-clustering of the gold-reagent. Underlying BCR oligomers of size 18 take over beginning from an observed gold particle size of 4. The contribution by BCR monomers is negligible. When testing a model with several BCR oligomer sizes, e.g., $1 + 2 + 18$, the maximization of the log-likelihood gives an estimate for the ratio between the abundance of the single BCR oligomer sizes. This ratio is then expressed in the percentage of receptors being contained in either of the oligomers. The number of receptors in oligomers of each size, computed from the model, indicates that 60% of all receptors are contained in BCR oligomers of size 18, 40% are contained in BCR dimers and BCR monomers are negligible (**Figures 4E,F**). In conclusion, we suggest that IgD-BCRs are arranged in pre-formed BCR oligomers on the surface of J558Lδm/mb-1flN B cells and that BCR dimers as well as large BCR oligomers of a laminar geometry co-exist.

3.6. DETERMINATION OF THE SIZE DISTRIBUTION OF A MUTANT IgD-BCR THAT MOSTLY FORMS DIMERS

To validate our combined experimental and theoretical approach, we took advantage of a transmembrane mutant IgD-BCR (IgD-hSbap BCR), which resulted in impeded BCR oligomerization as detected by Blue Native gel electrophoresis and pre-dominant detection of BCR dimers (16). Immuno-gold-labeling of the mutant BCR and sampling were performed as above. The quantification of gold clusters by TEM revealed mainly gold monomers (83%) and dimers (13%), and a smaller fraction of gold oligomers of three or more gold particles (4%) (**Figures 5A,B**). In contrast to the wt IgD-BCR, the staining efficiency was not determined as mutant IgD-BCR expression was heterogeneous (**Figure 3B**). When evaluating the log-likelihood for BCR staining efficiencies up to 40% and dominant BCR oligomer sizes up to10, the overall best log-likelihood value was achieved for BCR dimers at a

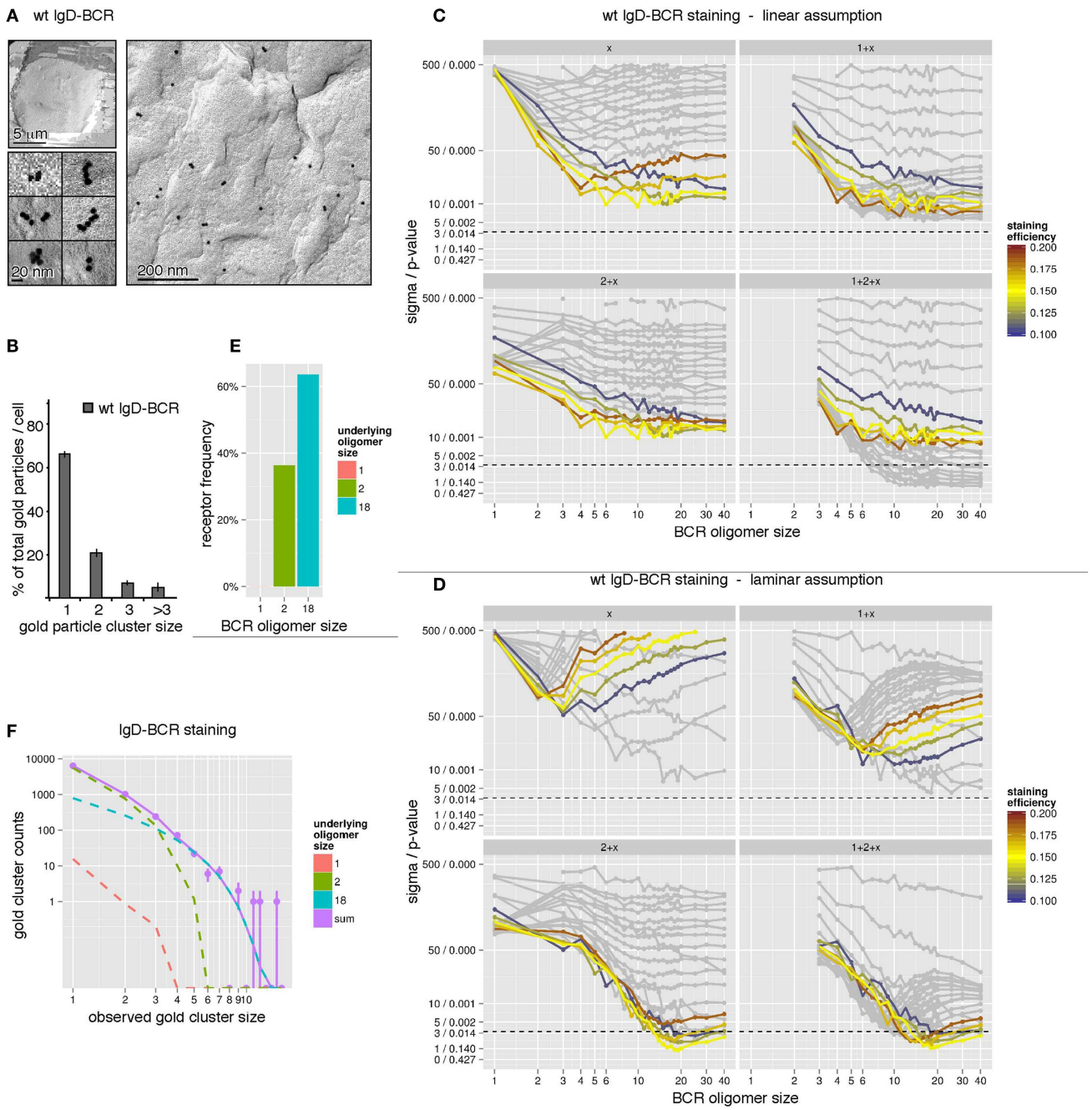


FIGURE 4 | The IgD-BCR forms differently sized oligomers on the surface of B cells. **(A)** IgD-BCR-expressing J558Lm/mb-1flN cells were immuno-gold stained using the mouse anti-BCR antibody Ac146 and anti-mouse IgG antibodies coupled to gold particles of 10 nm. The cell overview is given at low magnification (upper left panel), gold-labeling of the cell surface is shown in an overview picture (right panel) and individual gold clusters are shown at high magnification (lower left panel). **(B)** 66% of the gold particles were present as monomers, 21% formed dimers, 7.5% trimers, and 5.5% were present in clusters of sizes larger than 3. **(C)** The observed gold particle cluster size distribution was simulated with the assumption of an underlying linear BCR oligomer geometry. Simulations were performed for staining efficiencies of $15 \pm 5\%$ indicated by colored lines. The BCR oligomer sizes tested are indicated on the *x*-axis, and the sigma/*p*-values for the statistical analysis of each simulation set is given on the *y*-axis. Different hypothetical BCR oligomer size distributions are tested, namely "one BCR oligomer size only" (*x*, upper left panel), "BCR monomers + another single BCR oligomer size" ($1 + x$, upper right panel), "BCR dimers + another single BCR oligomer size" ($2 + x$, lower left panel), and "BCR monomers, dimers + another single BCR oligomer size" ($1 + 2 + x$, lower right panel). Models are rejected above the dashed line corresponding to a *p*-value of 0.01. **(D)** The observed gold particle cluster size distribution was simulated as in (C) with the assumption of an underlying laminar BCR oligomer geometry. **(E)** Our simulations predict that 60% of all receptors are part of oligomers of size 18 in a laminar geometry and 40% are BCR dimers. **(F)** Gold cluster counts (dots with error bars) compared to the prediction of best fitting model (continuous line). The number of counts originating from the different underlying BCR oligomer sizes are shown as dashed lines, indicating that observed clusters up to a size of 3 are primarily caused by BCR dimers, while gold clusters larger than 3 can be explained by BCR clusters of size 18. The impact of BCR monomers is negligible.

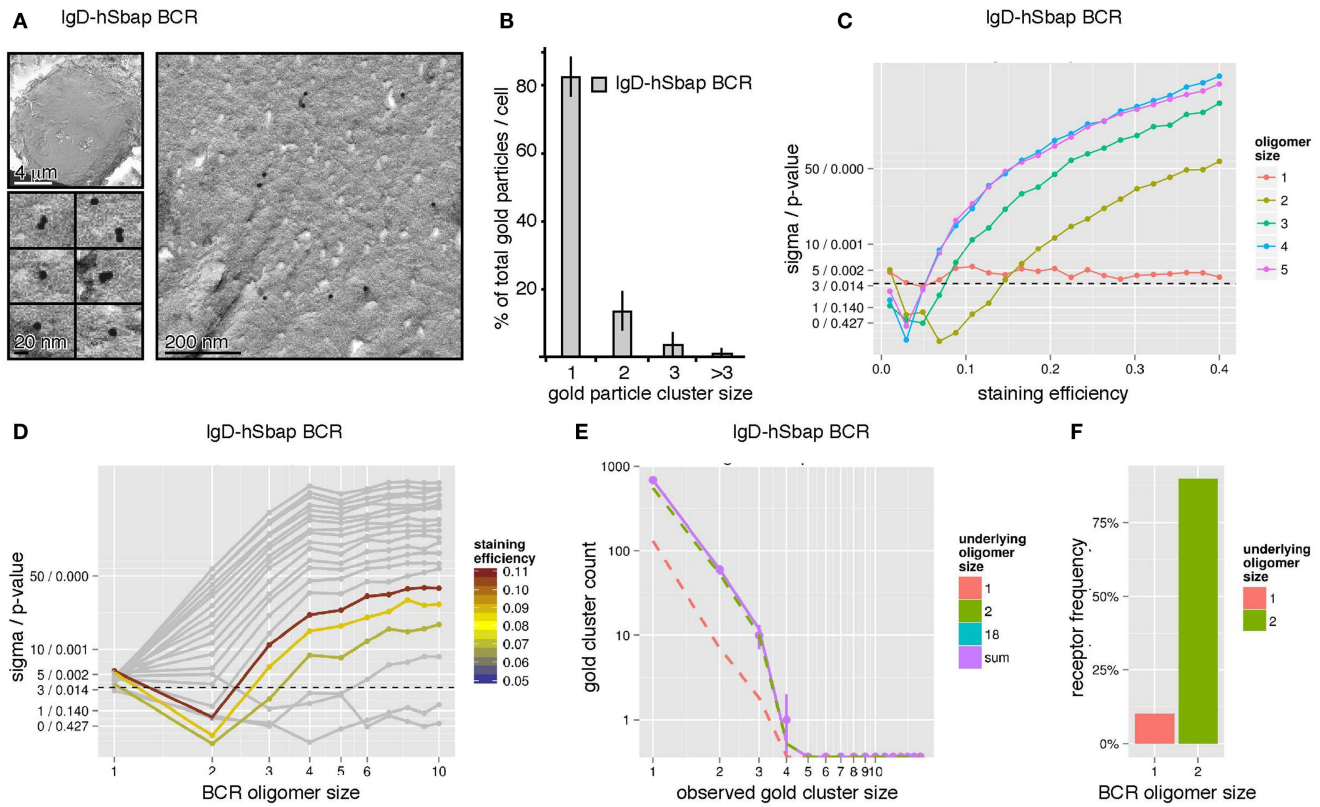


FIGURE 5 | The IgD-hSbap BCR mutant forms mainly BCR dimers on the surface of B cells. **(A)** IgD-hSbap BCR mutant-expressing J558L cells were immuno-gold stained and analyzed as in **Figure 4A**. **(B)** 83% of the gold particles were monomeric, 13% formed dimers, and 4% were present in clusters of sizes larger than 2. **(C)** The observed gold particle cluster size distribution was simulated for different BCR staining efficiencies (*x*-axis) and different underlying BCR oligomer sizes (color scale). The sigma/*p*-values for the statistical analysis of each simulation set are given on the *y*-axis. BCR dimers with a staining efficiency of 8% are identified as the most likely explanation for the observed gold cluster distribution. **(D)** Simulations were performed as in **(C)**. Assuming a staining efficiency of approximately 8% (color coding), sigma/*p*-values (*y*-axis) are plotted for the different BCR oligomer sizes (*x*-axis). **(E)** Gold cluster counts (dots with error bars) compared to prediction of best fitting model (continuous line). The number of counts originating from monomers and dimers are shown as dashed lines, indicating that the observed distribution is dominated by BCR dimers. **(F)** A model with BCR monomers and dimers was fitted to the data, predicting that 90% of all receptors are part of dimers, 10% are monomers.

staining efficiency of 8.8% (**Figure 5C**). The efficiency value is not too far from the wt IgD-BCR staining efficiency, which has been measured to be 15%. Our Monte-Carlo simulation indicated that IgD-hSbap BCR oligomers larger than two can be rejected at high confidence level unless the staining efficiency would be lower than 5% (**Figure 5D**). Also BCR monomers without larger BCR oligomers cannot explain the data, thus, confirming that the mutant BCR dimerizes. In addition to the single size model, a model with monomers and one further BCR oligomer size was tested. The observed gold cluster distribution after immuno-gold-labeling of the mutant IgD-BCR together with the model prediction of the best fitting model, i.e., monomers plus dimers, is shown in **Figure 5E**. Analogously to the wt IgD-BCR experiment, the percentage of IgD-hSbap BCR present as BCR monomers and BCR dimers was computed, demonstrating that 90% of the mutant BCRs are present as dimers and 10% as monomers (**Figure 5F**). These findings validate our approach and reflect the Blue Native gel electrophoresis data for IgD-hSbap BCR (16).

4. CONCLUSION

Here, we used a combined immuno-gold TEM and modeling approach to suggest that the IgD-BCR is expressed as BCR dimers and larger oligomers. To measure the distribution of cell surface molecules super resolution light microscopy techniques, such as high-speed photoactivated localization microscopy (5) or dSTORM (21), have been employed. Unlike these light microscopy techniques, our technique permits nano-scale resolution of cell surface molecules on the membranes that are not in contact with any support, thereby avoiding visualization of potential clustering artifacts induced by cell adherence (26). This advantage is not given in other EM methods, such as nano-probe labeling and transmission microscopy of cytoplasmic face-up sheets of cell membrane ripped off from cells to generate cytoplasmic face-up membrane sheets (27), which subsequently can be fixed and immuno-gold labeled (5, 28, 29). Our method can measure the distribution of any cell surface molecule, without the need to genetically modify this protein with a fluorophore, provided that antibodies against the protein are available. It is, however, not compatible with dynamic

observations and, given the relatively low labeling efficiency, does not allow absolute quantifications by direct gold counting; here we present an approach to overcome the latter limitation. We assumed that the presence of one gold particle is indicative of one BCR (with the restraint that some gold particles are aggregated, which we have taken into account). There is the unlikely possibility that our staining approach with a primary anti-BCR and a secondary gold-coupled antibody could allow for double labelings of the BCRs, if the secondary antibody binds twice to the primary antibody (once to each of the two heavy chains). However, double labeling most likely did not take place: we used the IgD-hSbap BCR that only formed monomers and dimers when analyzed by Blue Native gels (16). And indeed, in our analysis this mutant BCR was exclusively detected as monomers (10% of the BCRs) and dimers (90% of the BCRs). These relative abundancies fit well to the biochemical Blue Native gel data (16). Thus, we conclude that a potential double gold-labeling of a BCR does not take place in our experiments and does not prevent drawing our conclusions from the immuno-gold TEM data. Here, we estimate that 60% of all wild type IgD-BCRs expressed on J558Lδm/mb-1flN cells are part of BCR oligomers of a size of around 18 and 40% form BCR dimers. This notion is based on the most simple BCR size distribution (i.e., assuming the lowest number of different BCR sizes) to explain the observed gold distribution. Still, the co-existence of larger number of different BCR oligomer sizes cannot be ruled out. Previously, the oligomer size of the IgM-BCR has been theoretically approached by Iber and Gruhn (30). Assuming oligomer formation and decay rates based on the literature, they calculated that IgM-BCR pentamers might exist. However, the modeling approach was not based on experimental data on BCR oligomer sizes and thus the presence of BCR pentamers is rather hypothetical. BCR pre-clustering might increase the antigen-receptor avidity, since functional antigens are mostly multimeric structures (31, 32). This increase in avidity could enhance the sensitivity of B cell activation, as suggested for TCR pre-clusters (8–10). BCR oligomers could also enhance the sensitivity of B cells by cooperativity between BCRs in one oligomer. Thus, antigen-binding to one or two BCRs could also lead to the activation of non-engaged BCRs, propagating the signal within one oligomer. In line with this, we had suggested earlier that the kinase Syk bound to one BCR can phosphorylate the neighboring BCR amplifying BCR signaling by a positive feedback loop (33). In T cells it was shown experimentally that individual TCRs cooperate in this manner within one TCR pre-cluster (34). On the other hand, enhanced sensitivity of the B cells might also bear an increased risk of reacting against abundant self-antigens and lead to autoimmunity. The presence of pre-formed BCR oligomers has important implications for the mechanism of how the BCR transmits the signal of antigen-binding into the interior of the cells. Initially, it was assumed that the BCRs are individually distributed on the B cell surface and only brought into close proximity by their multivalent antigens, leading to reciprocal phosphorylation of the cytoplasmic tyrosines in Igα and Igβ and thereby signal initiation. The idea of cross-linking as the first step in BCR activation was based on the finding that monovalent antigens do not induce BCR activation (31, 32). Now, the cross-linking model is questioned by the finding that BCR oligomers exist (16, 17, 19, 21, 35). In our analysis we hardly detected any BCR monomers, which is in line with the finding that only oligomerized are stably expressed on the cell surface (19). Our study clearly points out that BCR oligomers of different sizes co-exist on the surface of resting B cells. Activation therefore might occur according to conformational changes within the BCR oligomer, as e.g., the proposed in the dissociation activation model (20). Knowledge on the pre-clustering of a given receptor not only aids to explain the biology of the receptor, but also might open new strategies for interfering with its function for vaccination or therapeutic purposes.

ACKNOWLEDGMENTS

We thank Balbino Alarcón, Hisse M. van Santen, and Maite T. Rejas from Madrid for their help in the preparation of the electron microscopy samples. This work was funded by the Excellence Initiative of the German Research Foundation (EXC294, BIOSS Centre for Biological Signalling Studies, and GSC-4, Spemann Graduate School of Biology and Medicine). The article processing charge was funded by the open access publication fund of the Albert Ludwigs University Freiburg.

REFERENCES

1. Singer SJ, Nicolson GL. The fluid mosaic model of the structure of cell membranes. *Science* (1972) **175**:720–31. doi:10.1126/science.175.4023.720
2. Simons K, Ikonen E. Functional rafts in cell membranes. *Nature* (1997) **387**:569–72. doi:10.1038/42408
3. Lingwood D, Simons K. Lipid rafts as a membrane-organizing principle. *Science* (2010) **327**:46–50. doi:10.1126/science.1174621
4. Alarcon B, Swamy M, van Santen HM, Schamel WWA. T-cell antigen-receptor stoichiometry: pre-clustering for sensitivity. *EMBO Rep* (2006) **7**:490–5. doi:10.1038/sj.embor.7400682
5. Lillemeier BF, Mortelmaier MA, Forstner MB, Huppa JB, Groves JT, Davis MM. TCR and Lat are expressed on separate protein islands on T cell membranes and concatenate during activation. *Nat Immunol* (2010) **11**:90–6. doi:10.1038/ni.1832
6. Schamel WW, Arechaga I, Risueno RM, van Santen HM, Cabezas P, Risco C, et al. Coexistence of multivalent and monovalent TCRs explains high sensitivity and wide range of response. *J Exp Med* (2005) **202**:493–503. doi:10.1084/jem.20042155
7. Sherman E, Barr V, Manley S, Patterson G, Balagopalan L, Akpan I, et al. Functional nanoscale organization of signaling molecules downstream of the T cell antigen receptor. *Immunity* (2011) **35**:705–20. doi:10.1016/j.immuni.2011.10.004
8. Kumar R, Ferez M, Swamy M, Arechaga I, Rejas MT, Valpuesta JM, et al. Increased sensitivity of antigen-experienced T cells through the enrichment of oligomeric T cell receptor complexes. *Immunity* (2011) **35**:375–87. doi:10.1016/j.immuni.2011.08.010
9. Molnar E, Deswal S, Schamel WW. Pre-clustered TCR complexes. *FEBS Lett* (2010) **584**:4832–7. doi:10.1016/j.febslet.2010.09.004
10. Molnar E, Swamy M, Holzer M, Beck-Garcia K, Worch R, Thiele C, et al. Cholesterol and sphingomyelin drive ligand-independent T-cell antigen receptor nanoclustering. *J Biol Chem* (2012) **287**:42664–74. doi:10.1074/jbc.M112.386045
11. Hombach J, Tsubata T, Leclercq L, Stappert H, Reth M. Molecular-components of the B-cell antigen receptor complex of the IgM class. *Nature* (1990) **343**:760–2. doi:10.1038/343760a0
12. Havran WL, DiGiusto DL, Cambier JC. mIgM:mIgD ratios on B cells: mean mIgD expression exceeds mIgM by 10-fold on most splenic B cells. *J Immunol* (1984) **132**:1712–6.
13. Scher I, Titus JA, Finkelman FD. The ontogeny and distribution of B cells in normal and mutant immune-defective CBA/N mice: two-parameter analysis of surface IgM and IgD. *J Immunol* (1983) **130**:619–25.
14. Reth M. Antigen receptor tail clue. *Nature* (1989) **338**:383–4. doi:10.1038/338383b0
15. Reth M, Wienands J, Schamel WW. An unsolved problem of the clonal selection theory and the model of an oligomeric B-cell antigen receptor. *Immunol Rev* (2000) **176**:10–8. doi:10.1034/j.1600-065X.2000.00610.x

16. Schamel WW, Reth M. Monomeric and oligomeric complexes of the B cell antigen receptor. *Immunity* (2000) **13**:5–14. doi:10.1016/S1074-7613(00)00003-0
17. Reth M. Oligomeric antigen receptors: a new view on signaling for the selection of lymphocytes. *Trends Immunol* (2001) **22**:356–60. doi:10.1016/S1471-4906(01)01964-0
18. Tolar P, Sohn HW, Pierce SK. The initiation of antigen-induced B cell antigen receptor signaling viewed in living cells by fluorescence resonance energy transfer. *Nat Immunol* (2005) **6**:1168–76. doi:10.1038/ni1262
19. Yang J, Reth M. Oligomeric organization of the B-cell antigen receptor on resting cells. *Nature* (2010) **467**:465–9. doi:10.1038/nature09357
20. Yang J, Reth M. The dissociation activation model of B cell antigen receptor triggering. *FEBS Lett* (2010) **584**:4872–7. doi:10.1016/j.febslet.2010.09.045
21. Mattila PK, Feest C, Depoil D, Treanor B, Montaner B, Otipoby KL, et al. The actin and tetraspanin networks organize receptor nanoclusters to regulate B cell receptor-mediated signaling. *Immunity* (2013) **38**:461–74. doi:10.1016/j.immuni.2012.11.019
22. Reth M, Hammerling GJ, Rajewsky K. Analysis of the repertoire of anti-NP antibodies in C57BL/6 mice by cell fusion. I. Characterization of antibody families in the primary and hyperimmune response. *Eur J Immunol* (1978) **8**:393–400. doi:10.1002/eji.1830080605
23. Fiala GJ, Rejas MT, Schamel WW, van Santen HM. Visualization of TCR nanoclusters via immunogold labeling, freeze-etching and surface replication. *Methods Cell Biol* (2013) **117**:391–410. doi:10.1016/B978-0-12-408143-7.00021-9
24. Murphy RM, Slayter H, Schurtenberger P, Chamberlin RA, Colton CK, Yarmush ML. Size and structure of antigen-antibody complexes. *Biophys J* (1988) **54**:45–56. doi:10.1016/S0006-3495(88)82929-1
25. Morath V, Keuper M, Rodriguez-Franco M, Deswal S, Fiala G, Blumenthal B, et al. Semi-automatic determination of cell surface areas used in systems biology. *Front Biosci (Elite Ed)* (2013) **5**:533–45. doi:10.2741/E635
26. James JR, McColl J, Oliveira MI, Dunne PD, Huang E, Jansson A, et al. The T cell receptor triggering apparatus is composed of monovalent or monomeric proteins. *J Biol Chem* (2011) **286**:31993–2001. doi:10.1074/jbc.M111.219212
27. Sanan DA, Anderson RG. Simultaneous visualization of LDL receptor distribution and clathrin lattices on membranes torn from the upper surface of cultured cells. *J Histochem Cytochem* (1991) **39**:1017–24. doi:10.1177/39.8.1906908
28. Wilson BS, Pfeiffer JR, Oliver JM. Observing FcepsilonRI signaling from the inside of the mast cell membrane. *J Cell Biol* (2000) **149**:1131–42. doi:10.1083/jcb.149.5.1131
29. Zhang J, Leiderman K, Pfeiffer JR, Wilson BS, Oliver JM, Steinberg SL. Characterizing the topography of membrane receptors and signaling molecules from spatial patterns obtained using nanometer-scale electron-dense probes and electron microscopy. *Micron* (2006) **37**:14–34. doi:10.1016/j.micron.2005.03.014
30. Iber D, Gruhn T. Organisation of B-cell receptors on the cell membrane. *Syst Biol (Stevenage)* (2006) **153**:401–4. doi:10.1049/ip-syb:20060015
31. Dintzis HM, Dintzis RZ, Vogelstein B. Molecular determinants of immunogenicity: the immunon model of immune response. *Proc Natl Acad Sci U S A* (1976) **73**:3671–5. doi:10.1073/pnas.73.10.3671
32. Minguet S, Dopfer EP, Schamel WW. Low-valency, but not monovalent, antigens trigger the B-cell antigen receptor (BCR). *Int Immunol* (2010) **22**:205–12. doi:10.1093/intimm/dxp129
33. Rolli V, Gallwitz M, Wossning T, Flemming A, Schamel WW, Zrn C, et al. Amplification of B cell antigen receptor signaling by a Syk/ITAM positive feedback loop. *Mol Cell* (2002) **10**:1057–69. doi:10.1016/S1097-2765(02)00739-6
34. Martinez-Martin N, Risueno RM, Morreale A, Zaldivar I, Fernandez-Arenas E, Herranz F, et al. Cooperativity between T cell receptor complexes revealed by conformational mutants of CD3epsilon. *Sci Signal* (2009) **2**:ra43. doi:10.1126/scisignal.2000402
35. Treanor B, Depoil D, Gonzalez-Granja A, Barral P, Weber M, Dushek O, et al. The membrane skeleton controls diffusion dynamics and signaling through the B cell receptor. *Immunity* (2010) **32**:187–99. doi:10.1016/j.immuni.2009.12.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; paper pending published: 28 August 2013; accepted: 20 November 2013; published online: 06 December 2013.
*Citation: Fiala GJ, Kaschek D, Blumenthal B, Reth M, Timmer J and Schamel WWA (2013) Pre-clustering of the B cell antigen receptor demonstrated by mathematically extended electron microscopy. Front. Immunol. **4**:427. doi: 10.3389/fimmu.2013.00427 This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.*
Copyright © 2013 Fiala, Kaschek, Blumenthal, Reth, Timmer and Schamel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 02 September 2013
doi: 10.3389/fimmu.2013.00263

The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor

Katherine J. L. Jackson*, Marie J. Kidd, Yan Wang and Andrew M. Collins

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Gur Yaari, Yale University, USA
Nir Friedman, Weizmann Institute of Science, Israel

****Correspondence:***
Katherine J. L. Jackson, School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: katherine.jackson@unsw.edu.au

Both the B cell receptor (BCR) and the T cell receptor (TCR) repertoires are generated through essentially identical processes of V(D)J recombination, exonuclease trimming of germline genes, and the random addition of non-template encoded nucleotides. The naïve TCR repertoire is constrained by thymic selection, and TCR repertoire studies have therefore focused strongly on the diversity of MHC-binding complementarity determining region (CDR) CDR3. The process of somatic point mutations has given B cell studies a major focus on variable (IGHV, IGLV, and IGKV) genes. This in turn has influenced how both the naïve and memory BCR repertoires have been studied. Diversity (D) genes are also more easily identified in BCR VDJ rearrangements than in TCR VDJ rearrangements, and this has allowed the processes and elements that contribute to the incredible diversity of the immunoglobulin heavy chain CDR3 to be analyzed in detail. This diversity can be contrasted with that of the light chain where a small number of polypeptide sequences dominate the repertoire. Biases in the use of different germline genes, in gene processing, and in the addition of non-template encoded nucleotides appear to be intrinsic to the recombination process, imparting "shape" to the repertoire of rearranged genes as a result of differences spanning many orders of magnitude in the probabilities that different BCRs will be generated. This may function to increase the precursor frequency of naïve B cells with important specificities, and the likely emergence of such B cell lineages upon antigen exposure is discussed with reference to public and private T cell clonotypes.

Keywords: BCR repertoire, TCR repertoire, V(D)J recombination, public clonotypes, private clonotypes, combinatorial diversity, junctional diversity

GERMLINE GENES AND LYMPHOCYTE DIVERSITY

The mammalian immune system has the ability to respond to almost any antigen to which it is exposed because of the incredible diversity of lymphocyte receptor molecules. The diversity of both the B cell receptor (BCR) repertoire and the T cell receptor (TCR) repertoire is made possible by multiple sets of highly similar genes that recombine to form functional genes. Immunoglobulin heavy chains are encoded by recombined VDJ genes that are formed from sets of Variable (V), Diversity (D), and Joining (J) genes (IGHV, IGHJ, IGHD), while VJ rearrangements of kappa and lambda chain V genes (IGKV, IGLV) and J genes (IGKJ, IGLJ) encode the immunoglobulin light chains (1, 2). TCR β-chains and δ-chains are similarly encoded by distinct sets of V, D, and J genes (TRBV, TRBD, TRBJ; TRDV, TRDD, TRDJ), while α-chains and γ-chains are encoded by additional sets of V and J genes (TRAV, TRAJ; TRGV, TRGJ) (3–5). The resulting combinatorial diversity is expanded still further by junctional diversification arising from exonuclease trimming of the recombining gene ends and from the essentially random addition of nucleotides, between the recombining genes, by the enzyme terminal deoxynucleotidyl transferase (TdT) (6). Together, combinatorial diversity and junctional diversity create the diversity of the naïve T cell and B cell repertoires. Limitations to diversity may however be a feature of V(D)J rearrangement that is as significant to immune function as the bewildering number of lymphocyte specificities that can theoretically be generated.

This review will present evidence that biases in the processes that generate combinatorial and junctional diversity are such that the probabilities of different BCRs and TCRs being generated is highly variable. This results in B and T cells of some specificities being present within the naïve repertoire at high frequency, while other specificities may or may not be present at all. The unevenness of the receptor abundance distribution can be said to give "shape" to the naïve B and T lymphocyte repertoires. This distribution may be further shaped by processes including positive and negative selection, clonal expansion and, in the case of immunoglobulin genes, by somatic hypermutation, however this review will focus upon recombination and gene processing.

As the shape of the naïve human B and T cell lymphocyte repertoire is an outcome of the evolution of genetically determined biases, this should ensure the presence of critical rearrangements in the repertoire of all individuals. It should also ensure that these critical rearrangements are carried by multiple naïve cells (see **Figure 1**). Such populations of specific naïve lymphocytes will have a competitive advantage during antigen-driven clonal selection, and any discussion of repertoire diversity that is limited to the size of the population of unique receptors will therefore be ignoring a parameter of likely biological significance. In this review, we

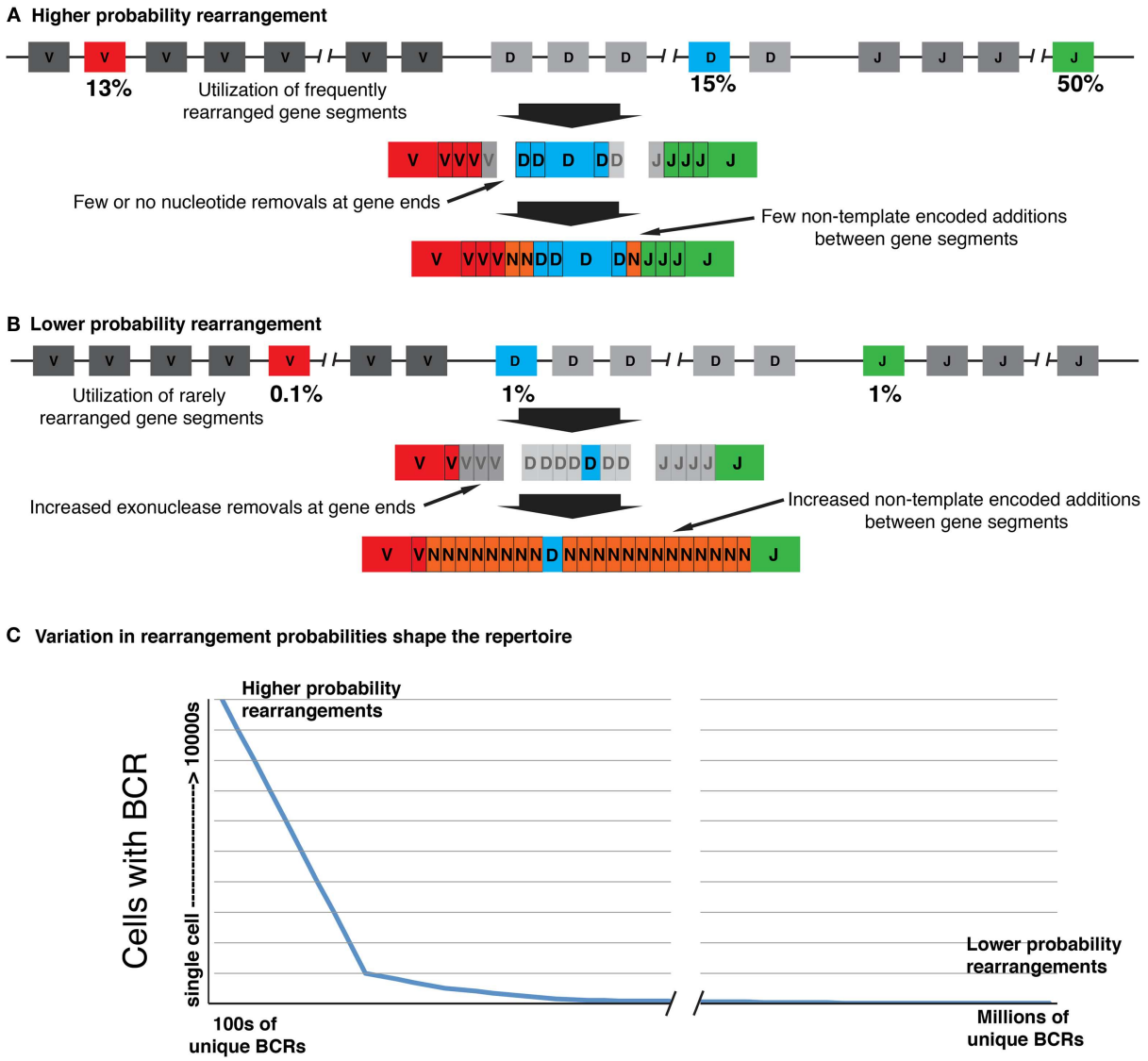


FIGURE 1 | The receptor repertoire has "shape," as biases and constraints in the recombination process vary the probabilities of generating particular V(D)J rearrangements. **(A)** Higher probability rearrangements are generated through utilization of more frequently rearranged gene segments. These segments are joined with minimal gene processing and N-addition, increasing the chance of independent rearrangements with identical or near identical CDR3s. **(B)** Lower probability rearrangements utilize rarely rearranged germline gene segments and the CDR3s are more diverse owing to increased nucleotide removals and additions at the joins. **(C)** The many order of magnitude differences in the likelihood of generating particular rearrangements shape the repertoire, with higher probability rearrangements being frequently generated and as a consequence being carried by many identical B-cells. Only a relatively small number of unique rearrangements will be generated with probabilities high enough to be carried by a large number of B cells, but this should ensure that they are always present in the repertoire at significant levels. Conversely, lower probability rearrangements may be so rare that they are carried only by a single B cell, or are entirely absent from the repertoire. The lower probability rearrangements that are carried by just one or at most a few B-cells likely represent many millions of unique rearrangements.

will use the term "repertoire" to refer to the complete set of receptors that are carried by an individual, including multiple copies of particular sequences. The number of unique sequences that are found within an individual's repertoire will be described as the "diversity" of the repertoire.

The size of the sets of germline genes make a major contribution to lymphocyte diversity, but surprisingly, our knowledge of these germline genes is far from complete. In part this is the result of the complexity of the loci, for they feature numerous highly similar genes that are thought to have evolved via gene conversion (7), and duplication and divergence (8). These genes are interspersed with many pseudogenes and repetitive elements (8). Sequencing and annotation of the loci is therefore challenging. These complexities also mean that SNPs arising from short read-length sequences generated in studies such as the HapMap and 1000 Genomes projects, cannot be used for the imputation of

full-length allelic variants. In fact, these projects utilize polyclonal lymphoblastoid cells lines in which the immunoglobulin loci have undergone somatic recombination, and the rearranged genes may have been affected by somatic point mutation. This makes these cell lines unsuited to the study of immunoglobulin genes (9).

Arguably, it is the BCR germline genes that are best known, and paradoxically, this is because of their transformation through the process of somatic hypermutation, during an immune response. IGHV genes are by far the longest of the recombining IGH genes, and they are the principal targets of the mutational machinery (10, 11). Many studies of the immunogenetics of immunoglobulin have therefore concentrated upon the IGHV genes. As it is necessary to be certain of the germline origin of mutated sequences, if accurate studies of point mutations are to be conducted, the complete and accurate definition of the set of germline immunoglobulin IGHV genes and allelic variants has been and should remain a focus of research.

The official human IGHV germline gene dataset, curated by the ImMunoGeneTics (IMGT) group, includes 129 functional genes, open reading frames (ORF), and pseudogenes, as well as over 200 allelic variants (12). Interest in these germline genes has increased in recent years, resulting in 40 new allelic variants being reported since 2005 (13–17). Many additional IGHV allelic variants have also been identified in recent high-throughput sequencing studies, through analysis of cDNA-derived VDJ gene rearrangements (18, 19), but these have not been accepted as part of the official IGHV dataset. We have designated alleles identified in this way with unofficial allele names using an indicator ("p") of their "putative" nature (e.g., IGHV3-9*p03) (15), and these additional alleles can be found in the UNSWIg repertoire (http://www.ihmmune.unsw.edu.au/unswig.php).

The official human light chain V gene datasets appear to be relatively complete and accurate, though few allelic variants have been reported (20). Nevertheless these few variants appear to be of functional and clinical significance. For example, a variant kappa gene allele was identified within the Navajo population and has been reported to account for the susceptibility of this population to infections (21).

The human IGH germline genes receive continuing attention while the IMGT human TCR germline gene datasets have barely changed since the complete sequences of the TCR gene loci were first described (22, 23). The IMGT TRBV dataset includes 65 functional genes, ORFs and pseudogenes, and just 13 allelic variants, and no new TRBV sequence has been added to the dataset since the publication of the complete sequence of the TRB locus in 1996 (22). Only three TRAV/TRDV sequences (24) in the IMGT dataset are derived from studies published since the reporting of the complete sequence of the TRAV/TRDV locus (23), and some variants that were described soon afterward still remain officially unrecognized (25). The incomplete nature of the IMGT TRBV, and TRAV datasets in particular are clearly highlighted in the literature, for sequencing studies have reported many SNPs in the coding regions of these genes. Subramanyan and colleagues reported 279 SNPs in a study of 63 TRBV genes in 10 individuals from each of four human populations (26). Of these reported SNPs, 114 were located in coding regions of functional TRBV genes (26). A similar study of 57 TRAV/TRDV genes in the same 40 individuals resulted in the discovery of 284 SNPs, 51 of which encode amino acid changes in the coding regions of the gene sequences (27). The allelic variants associated with these TRAV/TRDV and TRBV SNPs have not been reported in the literature or in sequence databases, and they have not been incorporated into the official gene datasets. This is surprising because the SNPs were identified through amplification and sequencing of full-length genomic sequences. It is also unfortunate, for studies of TCR polymorphisms have shown that they can be of functional significance (28, 29).

The BCR and TCR D loci contribute differently to the generation of diversity, and the differences in the nature of the loci have influenced BCR and TCR research directions. The 27 human IGHD genes include 25 functional genes, 23 of which are unique (30). Although some IGHD genes, especially those of the IGHD1 gene family, are very similar, there is considerable sequence diversity amongst the genes. The lengths of the IGHD genes vary from 11 nucleotides to 37 nucleotides, and almost all of them are substantially longer than the TRBD and TRDD genes. This length and the IGHD gene variability have made improvement in the identification of IGHD genes within VDJ rearrangements a challenging but achievable research goal. Pursuit of this goal has driven the development of immunoglobulin gene alignment utilities including SODA2 (31), IgBLAST (32), and iHMMunealign (33). The objective measurement of the performance of these utilities is made difficult, however, by a lack of appropriate data sets. Ideally, performance would be measured using rearranged sequences of known composition. As such sets are unavailable, clonally related sequences can be used (32, 33). We have also compared the performance of different utilities using a set of long-read pyrosequenced (Roche 454) IGH rearrangements from an individual with a homozygous deletion of six IGHD genes (34). This test measures performance by the number of VDJ rearrangements in the dataset that are said to include the absent IGHD genes. Together these studies demonstrate that IGHD genes can now be identified with confidence, and as a consequence, analysis of the BCR heavy chain complementarity determining region (CDR) 3 can include detailed analysis of IGHD gene usage, gene processing, and N nucleotide addition.

Analysis of the TCR CDR3 is not so easy. The two human TRBD genes are both short (12 and 16 nucleotides) and highly similar at their 5′ ends (22). This makes their identification in VDJ rearrangements particularly difficult. The TRBD genes within a VDJ rearrangement are likely to be flanked by N-REGIONS of non-template encoded nucleotides. These nucleotides are introduced through the action of the TdT enzyme, which is biased to the addition of guanine (G) nucleotides (35) and to the addition of homopolymer tracts (36, 37). Distinguishing TRBD gene ends from G-rich N nucleotides is difficult because the TRBD genes are G-rich at both their 5′ and 3′ ends. A final complication is that the two alleles of TRBD2 differ by just a single nucleotide. This critical nucleotide is flanked on both sides, in both alleles, by GGG motifs. For these reasons, few TCR studies have included detailed analysis of TRBD genes and their processing, or of the N-REGIONS that can only be defined after the identification of a TRBD gene segment within the CDR3. Even the most recently developed TCR alignment utility excludes identification of TRBD genes from its output (38).

Analysis of the VDJ junction in TRD rearrangements is equally difficult. The three human TRDD genes are just 8, 9, and 13 nucleotides in length (4). This makes their reliable identification within VDJ rearrangements especially problematic if nucleotides have been lost through exonuclease activity. Application of an approach previously used in the analysis of BCR sequences (37) suggests that eight nucleotides is the minimum D gene length that will allow TRDD genes to be reliably distinguished from N-REGIONS within a junction of 12 or fewer nucleotides, while 9 nucleotides are needed for regions from 13 to 15 nucleotides and 10 nucleotides for junctions greater than 15 nucleotides (Jackson, unpublished data). It is therefore no surprise that few studies have reported the partitioning of TRD junctions as two of the three TRDD genes can only be confidently delineated from N-additions in their unprocessed form.

The J loci of the human BCR and TCR also include important differences. The IGHJ locus includes six functional genes, which are all found downstream of the IGHD locus in a single cluster. Allelic variants have been reported for IGHJ3, IGHJ4, IGHJ5, and IGHJ6, though there is reason to doubt the existence of the reported allelic variants of IGHJ3 and IGHJ5 (39). TCR J genes are more numerous and are differently organized. The TRBJ genes are found as a block of six genes located downstream from the TRDB1 gene, and a block of seven genes located downstream from the TRDB2 gene. The TRDB1 gene can pair with all J genes, but the TRDB2 gene is strongly biased toward pairing with its associated J genes (40). There are also four functional J genes in the TRDJ locus. Functional allelic variants have only been reported for the TRBJ1-6 gene.

BIASES IN COMBINATORIAL DIVERSITY AND THE SHAPING OF THE REPERTOIRE

Combinatorial diversity is that part of repertoire diversity that results from the fact that functional receptor genes form by the recombination of members of the sets of germline V, D, and J genes. This diversity is usually calculated by simply multiplying together the number of functional V, D, and J genes that are available within the genome. Such calculations, however, may promote misunderstandings, for they encourage the view that "all genes are equal," and that all combinations are equally likely. TCR studies have paid considerable attention to capturing an unbiased sampling of the repertoire, for example using 5′ RACE to amplify TCR transcripts from the constant region gene. Such studies have shown that TCR genes are highly biased in their usage (41–43). In contrast, many BCR repertoire studies have amplified both mRNA and genomic rearrangements, often using IGHV gene family-targeting primer sets that were developed for the detection of malignancies rather than for the investigation of the repertoire (44, 45). Such primers almost certainly lead to some distortions in the relative abundances of different sequences that are seen. Nevertheless, BCR studies utilizing different primer sets, and amplifying different source material are surprisingly consistent, and the B cell literature provides unequivocal evidence of strong gene utilization biases.

Different IGHV genes are used at frequencies that range from as little as 0.1% to more than 10% of all rearrangements in an individual's naïve B cell repertoire (18, 46). Utilization frequencies also vary between alleles. For example, analysis of VDJ recombination in different individuals has shown that IGHV1-2*02 is used approximately three times as often as IGHV1-2*04, in individuals who carry both these alleles (18). IGHV utilization frequencies are surprisingly constant between individuals (47). Examples of such consistency include IGHV1-46 which varies from 2 to 3.1% in different individuals (average 2.65%), IGHV3-21 which varies from 3.5 to 6.3% (average 4.59%), and IGHV3-49 which varies from 0.8 to 1.3% (average 1.0%) (18). This is not true for all genes, with different individuals utilizing IGHV1-69 at frequencies that range from 3.1 to 9.1% (average 6.2%) (18). IGHV3-23, which is typically the most utilized IGHV gene, was seen on average in 6.7% of all VDJ sequences, but its utilization frequency in one individual was 13.7% (18).

Biased gene usage is not confined to the IGHV genes. IGHD gene usage varies from less than 1% (IGHD4-4/11) to over 15% (IGHD3-22) of total rearrangements. Biases in the resulting amino acid sequences of the CDR3 junction are even greater. IGHD segments can be utilized in all three reading frames, and each IGHD gene is therefore able to encode three distinct amino acid sequences. Analysis of IGH rearrangements in which the IGHJ is out-of-frame, and which are therefore non-productive, shows each IGHD gene rearranges at equal frequency in each of the three RFs, however among productive rearrangements there is a strong skewing of the utilization of each gene toward a dominant RF (48). This dominance is constant between individuals, and the preferred RF is gene family dependent. Analysis of in-frame and out-of-frame IGH rearrangements sequenced using the Illumina platform suggests that the underlying rearrangement processes have no reading frame bias, but that bias emerges from stronger negative selection of sequences in certain reading frames (48). Such negative selection particularly focuses on non-productive sequences that result from the presence of stop codons within the junction region. These are seen when many IGHD genes are translated in the non-dominant reading frame, and such genes can only be utilized in those reading frames if the stop codons are removed by exonuclease trimming. When analysis of IGHD usage in the expressed repertoire factors in the three RFs, the IGHD gene utilization frequencies span three orders of magnitude. There is also considerable variation between the utilization frequencies of IGHJ genes. The IGHJ4 gene is present in approximately 45–50% of rearrangements, while IGHJ6 accounts for a further 20–25% of VDJ rearrangements (49, 50). IGHJ1, on the other hand, is utilized by only 1% of all rearrangements (39).

Biases in light chain gene usage are just as strong. For IGK rearrangements, preferential inclusion of IGKV3-20 was noted in early studies of the expressed IGK repertoire of both adults and neonates (51–53), while single cell PCR (54) and bioinformatics analysis of IGK rearrangements from sequence databases showed IGKV3-15, IGKV3-11, IGKV1-5, IGKV2-30, and IGKV1-30/IGKV1D-39 to also display preferential rearrangement (20). These biases were confirmed again recently in a high-throughput sequencing study which also highlighted similarities in usage between individuals, including similarities between individuals from geographically distant and ethnically distinct populations (55). Under-utilization and over-utilization of the J gene segments have been reported. IGKJ1 and IGKJ2 appear more frequently,

while there is under-utilization of IGKJ3 and IGKJ5 (20, 53, 54). This skewing of IGKJ usage toward the genes located 5′ in the IGKJ locus is seen despite the necessity for selection of more 3′ IGKJ genes during secondary IGK rearrangements (56, 57). A similar bias toward 5′ IGKJ genes is also seen in the mouse, and modeling of mouse light chain rearrangement supports the strong underlying tendency toward the initial rearrangement of IGKJ1 or IGKJ2 (58). The IGLV usage is strongly skewed toward a limited number of the functional V segments with 3 of the 30 IGLVs accounting for more than 50% of expressed rearrangements, and with individual IGLV segment frequencies ranging from 0.02 to 27% (59). Only four of the seven IGLJ are considered functional (60). The four IGLJ range from almost 55% utilization in the expressed B cell repertoire for IGLJ7, to just 5.5% for IGLJ1 (61).

Although bias in the reading frame of the IGHD gene is the result of selection, other biases appear to be intrinsic to the recombination process, for when analysis is confined to non-productive rearrangements which carry an out-of-frame J-REGION, preferential gene usage is still seen (48). Such sequences are not subject to positive or negative selection. The same biases have been observed among transcripts generated from transgenic mice that carry a human heavy chain mini-locus (62), while in NOD-scid-IL2Rγ null mice that had been reconstituted with human hematopoietic stem cells, typical patterns of biased usage were seen amongst the expressed light chain genes (63). Recent studies in monozygotic twins show that they share utilization frequencies for both the heavy and light chain genes (46, 63), with correlations in a similar range to replicate biological samples. When one twin was investigated following lymphocyte ablation therapy, the reconstituted repertoire showed the same utilization patterns (46). Unrelated individuals did not share this degree of correlation. The biases in utilization frequencies of different V, D, and J genes therefore appear to be genetically determined, and when acted upon by the recombination machinery, the biases in that process give rise to an individual's distinct repertoire. Repertoire shape is therefore directly linked to the genotype of an individual's immunoglobulin gene loci. This has become even clearer since high-throughput sequencing has allowed analysis to focus upon individual chromosomes.

The large datasets that are now being generated by high-throughput sequencing from single individuals are facilitating analysis of the processes that shape the repertoire, but each dataset still represents a mixture of rearrangements from two independently recombining chromosomes. The fact that V(D)J rearrangement is an intra-chromosomal event, however, means that every V(D)J gene rearrangement provides information about the association of different genes on a chromosome. Any heterozygous locus allows each chromosome to be associated with one or the other allele at that gene locus, and large sets of V(D)J rearrangements can be analyzed to determine all the V, D, and J genes that rearrange on each chromosome. This allows the determination of inferred haplotypes (see **Figure 2**).

In practice, the complete inference of V, D, and J gene haplotypes by the analysis of V(D)J rearrangements is only likely to be possible in the case of the IGH locus. Approximately 40% of individuals are heterozygous at the IGHJ6 locus, and the IGHJ6 gene is present in nearly 25% of all rearrangements. It therefore provides an ideal "anchor-point" from which to haplotype the IGH locus. Using this approach, we recently investigated the IGH locus in nine individuals, and showed that all 18 IGH variable region gene

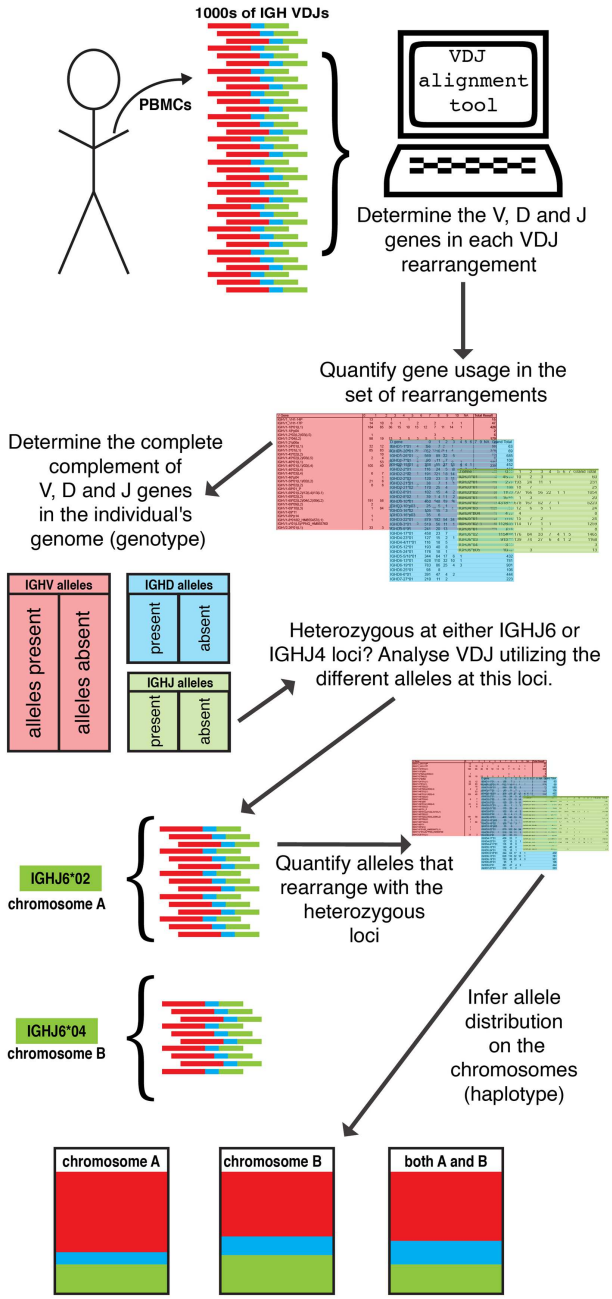


FIGURE 2 | Inference of IGH haplotypes from VDJ rearrangements. The availability of large datasets of VDJ rearrangements from single individuals permit the inference of all germline V, D, and J genes within the genome, from analysis of apparent gene utilization within a dataset. As IGH VDJ rearrangement is an intra-chromosomal event, gene pairing in VDJ rearrangements can be used to infer which gene segments are carried by the same chromosome. Leveraging the heterozygous IGHJ6 or IGHJ4 locus allows the reconstruction of IGH gene segments on each chromosome and for the IGH haplotype to be inferred.

haplotypes were unique (19). In addition to allelic variants, many IGHV and IGHD gene deletions and IGHV gene duplications were evident. The definition of haplotypes in this way is allowing IGH gene usage frequencies to be studied with unprecedented accuracy, but unfortunately no locus as appropriate as the IGHJ6 anchor-point exists amongst the light chain genes or amongst TCR genes. Limited investigations in the past have highlighted TCR haplotypic variation in the human population (64, 65), but the extent of variation within the IGH locus suggests that considerably more TCR variation may await discovery.

Many factors have been explored to explain biases in chromosomal recombination patterns. Variations in enhancers (66) have been implicated in biased murine TCR gene usage. Variations in recombination signal sequences (RSS) also influence utilization frequencies of both human BCR (67, 68) and TCR (69) genes. The IGKV polymorphism that has been linked to increased susceptibility to *Haemophilus influenzae* in the Navajo population includes a single nucleotide change in the heptamer sequence of the RSS, and it reduces recombination by 4.5-fold relative to the common allelic variant (21). The non-amer and heptamer sequences of the RSS are separated by either a 12 or 23 base pair spacer. Spacers also show sequence variation, and there has been debate about the impact this has on recombination efficiency. While some studies did not observe any impact when the regular spacer sequence was replaced with runs of GC pairs (70), competition assays using extra chromosomal substrates suggest differences in spacer sequence can result in differences in recombination efficiency that mirror differential gene usage in the V(D)J repertoire (67, 68). However, RSS variation cannot explain all differences in allele utilization. The recent re-sequencing of the complete IGH locus found that the IGHV-associated RSS were the same as those earlier reported by Matsuda (71) even where different alleles of the gene were present (17).

Some variation in the frequency with which particular gene sequences are seen in the repertoire may be explained by copy-number variations (CNV). The presence of CNV within the IG variable gene locus was first determined using sequence-specific RFLP analysis to determine gene copy-number (72), and the affect of CNV on expression levels was investigated through the examination of the binding of an anti-idiotypic monoclonal antibody (G6) to tonsillar IgD + B-cells (73). An examination of 35 individuals found that they carried between 0 and 4 copies of the IGHV1-69 gene. Linear regression determined that for each allele copy, approximately 3% of B-cells were G6 reactive. Individual differences in the IGHV1-69 copy-number could therefore result in the contribution that this single gene makes varying from being totally absent (0 copies) to being present in as many as 12% of rearrangements in individuals with four available copies.

Sequencing of single chromosomes of an individual's IGH locus has now demonstrated that insertions, deletions, and complex events have altered the copy-number of IGHV genes, including the IGHV1-69 and IGHV3-23 genes (17). The duplicate IGHV3-23 genes remain within the genome as absolutely identical sequences. The presence of these and other CNVs has also been highlighted in bioinformatic studies of immunoglobulin genotypes (18) and haplotypes (19), where sequence data from single individuals clearly demonstrated that some individuals had more than two "alleles" of a single IGHV gene. Genes were also found to be absent from the genome of some individuals. A limitation of these bioinformatics studies was that gene duplications could only be detected if two distinct "allelic variants" were carried on a single chromosome.

In addition to the underlying biases in utilization of germline genes, a final bias has been identified that affects the contribution of recombination frequencies to repertoire diversity. For reasons that are presently unclear, there appear to be pairing preferences for some IGHD and IGHJ genes that increase the frequency of particular IGHD-IGHJ pairs within the repertoire. Biases were first observed in a small set of 59 non-productive rearrangements (74). Later analysis of 6,500 IGH VDJ sequences collected from public databases led to the observation that 5′ IGHD genes paired with increased frequency to the most 3′ IGHJ (J5/J6) and with decreased frequency to the 5′ IGHJ (J1–J4) (50). In contrast, 3′ IGHD tended to preferentially pair with 5′ IGHJ rather than 3′ IGHJ (50). This observation is also supported by analysis of very large datasets generated by pyrosequencing of VDJ rearrangements from three healthy subjects (75). Significantly more pairings were seen of IGHD2-2 and IGHD3-3 with IGHJ6, and of IGHD3-22 and IGHJ3 than would be predicted from the frequencies of these genes in the overall dataset (75).

The bias in D-J pairing also extends to the TCRB loci where the application of HTS approaches to murine TCRB repertoires has revealed a pattern of TRBD to TRBJ pairing that correlates to the genomic distance between rearranged genes (40). The TRBV and TRBJ gene usage in the mice was biased toward particular genes, but the pairings of TRBV and TRBJ were independent. The physical chromatin structure of the TRBD and TRBJ loci was investigated using a biophysical model of the chromatin conformation. The biases in TRBD to TRBJ pairing appeared to be better explained by this mechanical model than previously proposed genetic models based on RSSs (40). The model was also extended to human TRBJ usage with favorable evidence that chromatin conformation determines TRJB gene usage.

Biases in the pairing of heavy and light chains have also been reported. The existence of forbidden or unfavorable pairings of germline heavy and light chain genes was described in the early literature (76). This was not supported by later studies (77, 78), nor was it supported by a recent study that applied high-throughput sequencing to generate thousands of linked heavy and light chain genes (79).

BIASES IN JUNCTIONAL DIVERSITY AND THE SHAPING OF THE REPERTOIRE

Both the naive B cell and T cell repertoires are limited in the periphery by processes of selection. However T cell selection within the thymus is a particularly rigorous process, and it leads to dramatic differences between the potential and the observed repertoire diversity. The idiosyncratic nature of TCR selection in a human population with abundant MHC diversity also means that analysis of the processes that contribute to TCR diversity will be difficult using datasets comprising sequences from multiple individuals. Sufficiently large datasets from single individuals with a specific MHC profile finally became available with the application of high-throughput sequencing to repertoire studies. However, the continuing difficulties involved with the identification of TCR

D genes and hence the other constituent elements within the TCR CDR3 still discourage analysis of the genetic elements and processes that contribute to this region. It is therefore studies of BCR genes that provide the clearest insights into the processes that contribute palindromic P nucleotides and non-template encoded N nucleotides to the V(D)J junction, and into the process of exonuclease trimming that depletes the ends of recombining genes. A recent study of BCR CDR3 suggested that as a result of these processes, the circulating B cell population in a typical adult human includes 3–9 $\times$ 10^6 unique heavy chain CDR3 (80).

Palindromic or P nucleotides are formed by the asymmetric opening of hairpin loops that form at gene ends during the rearrangement process (81). In the absence of exonuclease activity, the opening of the hairpins can add short, self-complementary single stranded extensions into the junctions. P nucleotide addition was first recognized as a process that can contribute to TCR CDR3 (82, 83), however the contributions of P nucleotides to the BCR repertoire have been more precisely quantified (84, 85). Similarly, it is recognized that N nucleotides make a major contribution to the diversity of both the TCR and BCR repertoire (86), but only BCR N-REGIONS have been subjected to detailed analysis (37). Where BCR studies have investigated the kinds of amino acids that are likely within N-REGIONS, studies of TCR N-REGIONS have focused upon analysis of the overall contribution of N-REGIONS to αβ TCR diversity. This has been studied in a comparison of wild-type mice with mice carrying homozygous null alleles for TdT (86). N-addition was estimated to contribute to 90% of the diversity of the αβ TCR repertoire (86). Diversity could be estimated in this and other studies because of the development of "spectratyping" techniques, which is the analysis of the CDR3 length distribution in PCR amplicons. It permitted some of the first explorations of the T cell repertoire, however it only allowed detailed analysis of N-REGIONS if further sequencing was undertaken. Until the advent of high-throughput sequencing, such analysis was usually compromised by the restricted number of sequences that could be generated from any individual, and by the challenges associated with D gene identification.

Non-template encoded N-additions are intrinsically biased owing to the preference of TdT toward the incorporation of G nucleotides. This is manifested in G/C-rich additions when viewing the N-REGIONS of the coding strand, as additions may be made to both the coding and non-coding strands during recombination. This has been demonstrated through analysis of extrachromosomal substrates transfected into human cell lines (36), as well as by analysis of human BCR (37) and TCR (87) VDJ rearrangements. The G/C bias is coupled with an apparent interdependence of the additions, which leads to the formation of homopolymer tracts (36, 37, 87). Together these biases ensure that the germline gene-encoded regions of the CDR3 are frequently flanked by amino acids such as glycine, that are encoded by G-rich codons (88). It has been proposed that the inclusion of small amino acids such as glycine, which has only a single side chain, promotes flexibility of the CDR3 loop (88).

Exonuclease trimming is perhaps the least understood process that contributes to the BCR and TCR repertoires. The mechanisms responsible for the loss of nucleotides from the coding ends of the genes during rearrangement remains to be determined, but a number of features of the process have been described, and intrinsic biases have been identified. The extent of processing from each gene end involved in a join (VD or DJ) is independent (87). That is, we do not see more processing on one side of the join to compensate for reduced processing of the gene on the other side. The processing differs for V, D, and J genes and for gene families. Removals may therefore be impacted by the sequence of the gene ends. Sequences with high A/T content appear more susceptible to nucleotide loss, while sequences with high G/C content appear resistant to processing (36, 84, 89, 90). This bias is still seen after controlling for the G/C bias of N-REGIONS.

The gene sequence ends that remain after exonuclease processing provide a final bias that shapes the repertoire. The gene ends are constrained by the genetic code, to favor the formation of codons for a surprisingly limited number of amino acids. This is best illustrated in the case of the many IGHD genes that have the nucleotide sequence TAC at their 3′ end (see **Figure 3**). In the dominant reading frame, these nucleotides encode tyrosine. Removal of a single nucleotide creates a situation where only provision of a T or C (from N-addition or from the 5′ end of the IGHJ gene) will result in a functional sequence, for TAA and TAG are stop codons. Addition of C returns the sequence to its original state, while addition of T results in an alternative tyrosine codon. In this and other cases, the nucleotide sequences of the gene ends limit the diversity that results from exonuclease removals.

B CELL LINEAGES AND T CELL CLONOTYPES IN THE ANTIGEN-SPECIFIC RESPONSE

Biases that we have described in immunoglobulin V, D, and J gene usage mean that at least seven orders of magnitude separate the probabilities that the most likely and the least likely combinations of recombining genes will be generated in the bone marrow. Many additional orders of magnitude separate the most likely from the

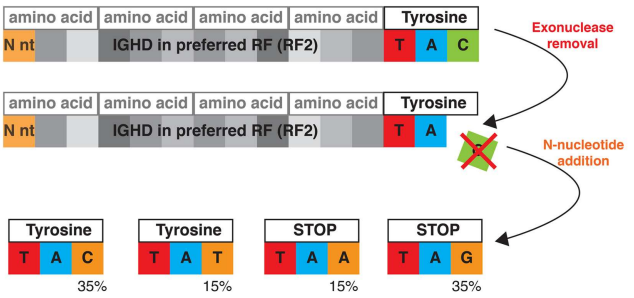


FIGURE 3 | Germline gene segment sequences constrain junctional diversity. This is illustrated by processing of the codon TAC, which encodes tyrosine as the most 3′ amino acid in the preferred reading frame of many IGHD genes. A highly probable outcome of the VDJ rearrangement process is the loss of a single nucleotide from the IGHD coding end. If this occurs, the productive repertoire for these Ds is strongly skewed toward tyrosine, as two of the four possible nucleotide replacement events generate stop codons, and the other two ensure the maintenance of tyrosine. These nucleotides may be sourced from either N-addition or from the 5′ end of an IGHJ segment. The outcome of N-addition is depicted in the bottom panel. N-additions are colored in orange and below each of the four possible outcomes is shown the likelihood of the enzyme TdT adding each nucleotide base.

least likely heavy and light chain pairs. The least likely BCRs are so unlikely to be generated in the bone marrow that they almost certainly will never be seen in an individual's lifetime. The most likely BCRs, on the other hand, may be so readily generated that they are always present within the repertoire at high copy number. These high copy-number sequences are likely to utilize a relative handful of the available germline genes, and to have been subject to minimal processing. TdT adds, on average, around 6 nucleotides between joining genes, and 30 or more nucleotides may occasionally be added to the VD, DJ, and VJ junctions, but it is highly likely that no more than two nucleotides will be added. Even long heavy chain CDR3 are likely to be the result of long germline sequences rather than the result of long N-REGIONS (47). Six or more nucleotides may be removed from the 3′ end of the IGHV gene, but most sequences lose no more than two or three IGHV nucleotides, and many sequences lose no nucleotides at all.

Without the added diversity that comes from D genes, the kappa and lambda repertoires are strongly shaped by biased gene usage and minimal processing and the diversity of the repertoires is surprisingly limited. The light chain repertoires are dominated by a very small number of amino acid sequences, and this dominance is so extreme that even in the days of Sanger sequencing, identical light chain gene rearrangements were reported by separate studies from independent laboratories (20). The theoretical diversity of the kappa repertoire has been estimated to be as high as 4×10^{24} unique nucleotide sequences (91). However analysis of kappa sequences generated from single individuals by high-throughput sequencing suggest the repertoire may include less than 10^4 unique amino acid sequences (55), and some of these sequences may be seen in over 1% of all kappa-bearing BCR (55). The diversity of the expressed lambda repertoire has recently been shown to be similarly restricted (63).

Although the heavy chain repertoire has much greater diversity then the light chain repertoire, repertoire shaping may be sufficiently extreme that some heavy chain sequences, and even some BCR will be present at high copy number in the repertoire of every individual. We are not aware of identical heavy chain sequences being amplified from multiple individuals, but highly similar "stereotypical" sequences have been found amongst leukemic clones of individuals with chronic lymphocytic leukemia (92). These stereotypical sequences differ through the stochastic processes of somatic hypermutation, but they appear to have evolved from cells expressing highly similar BCR within the naïve B cell repertoires of different individuals. Antigen selection, which may be associated with the pathogenesis of this condition (93), could be selecting and therefore revealing high copy-number heavy chain sequences.

The antigen specificity of most heavy and light chain sequences remain unclear, for it is only very recently that antigen-specific human B cells have been isolated and their BCRs investigated. The isolation of antigen-specific plasmablasts from the peripheral blood shortly after vaccination was first used to produce monoclonal human antibodies (94). These cells express BCR genes that are at once similar, as a consequence of their shared origins, yet highly divergent, as a result of the process of somatic point mutation. Together they make up a B cell clone lineage. High-throughput sequencing has since been used to identify clone lineages after booster shots with the influenza vaccine (95) and the pneumococcal vaccine (96). B cell lineages producing broadly neutralizing antibodies to HIV have also been identified using high-throughput sequencing (97). However this handful of studies of antigen-specific B cells in humans has not identified lineages that are shared between individuals. Highly similar BCR heavy chain sequences have recently been identified using high-throughput sequencing of PBMC from multiple individuals with acute symptomatic dengue (98). Although the specificities of these sequences were not determined, such lineages were not identified in uninfected individuals. These may therefore be the first antigen-specific heavy chain "public lineages" to be identified. The extent to which the response to specific antigen more generally involves such "public lineages" remains to be determined.

In contrast to the paucity of studies of antigen-specific B cells, antigen-specific TCRs have been investigated in the human repertoire for over 20 years. Early studies revealed that the immune response to specific antigen, in HLA-matched individuals, can include sets of T cells sharing identical or highly similar TCR α- and β-chains (99–101). The development of techniques for the creation of MHC peptide tetramer complexes has facilitated the identification of antigen-specific T cells by flow cytometry (102). This has allowed the detailed investigation of dominant sequence sets and these studies gave rise to the notions of public and private T cell "clonotypes." Public clonotypes are defined as VDJ amino acid sequences that are dominant and identical, or nearly identical, in multiple individuals. Private clonotypes, in contrast, are idiosyncratic. The apparently antigen-driven emergence of public B cell lineages in chronic lymphocytic B cell leukemia also has parallels amongst T cell leukemias. Studies of T cell large granular lymphocyte leukemias have identified a public clonotype in individuals with the shared DRB1*0701 HLA type (103). This same clonotype was independently identified in DRB1*0701$^+$ individuals who were infected with human cytomegalovirus (104), suggesting that antigen-driven pathogenesis may be expanding and revealing this public clonotype.

To understand the reasons for the emergence of particular clonotypes, the naïve repertoire must be better understood. Enrichment techniques have recently been developed which when combined with MHC peptide tetramer technology allows extremely rare peptide-specific naïve murine T cells to be identified (105). Using this approach in humans, naïve CD8$^+$ T cells specific for peptide-MHC have been shown to range from 0.6 to 500 cells per million cells (106, 107) and CD4+ T cells to range from 0.2 to 10 per million cells (107). Most of the cells within identified sets of antigen-specific murine T cells express unique TCRs (105), but clonal diversity within identified human cell populations remains unclear. It is likely though that in the much larger human T cell compartment, many circulating T cells could carry identical TCRs. This should ensure that early adaptive responses to these antigens are robust, for the strength of the response to antigen has been shown to reflect the size of the antigen-specific naïve T cell population (105).

The presence of particular public TCR clonotypes have not yet been reported within the naïve human TCR repertoire. Discussion of the emergence of such TCR clonotypes in an antigen-specific response has therefore been driven principally by analyses of

their nucleotide and amino acid features, and the phenomenon of convergent recombination has been invoked to explain public clonotypes (108, 109). Many public TCR clonotypes are divergent at the nucleotide level, but identical at the amino acid level. This results from the fact that particular amino acid sequences can arise from multiple, variant nucleotide sequences, and that these nucleotide sequences in turn can sometimes be formed by different genes with varying levels of gene processing and nucleotide addition. Such convergent recombination will certainly contribute to the presence of multiple copies of particular amino acid clonotypes within an individual's repertoire, but arguably, it is unlikely to increase the likelihood of one clonotype over another by more than one or two orders of magnitude.

More recently the role of biases in gene usage and in the recombination process have been identified as an alternative source of public clonotypes (110). The biases in the usage of TCR V, D, and J genes are less pronounced than is the case for the BCR genes. This is the result of the lack of substantial germline diversity within the sets of TRBD and TRDD genes, and because the TRBJ and TRDJ genes lack the strong usage biases that are seen amongst the IGHJ genes. Nevertheless biases in the usage of TCR genes are still likely to ensure that the probabilities of the generation of the least likely and the most likely V(D)J combination seen in αβ and γδ TCR differ by many orders of magnitude. It has also been pointed out that many public clonotypes have short CDR3 loops that are mainly encoded by germline-derived nucleotides rather then TdT-derived nucleotides (110). The contribution this may make to the formation of T cell clonotypes is harder to judge, because of the lack of detailed analysis of these processes, in the context of the TCR repertoire. However lessons from analysis of the BCR repertoire give strong credence to this hypothesis.

Both the BCR and the TCR repertoires have been the subject of considerable study and even greater speculation over many decades. High-throughput sequencing is now revealing their separate secrets at a gratifying rate. Our understanding of the shaping of the BCR and TCR repertoires will now surely move faster if a greater dialog commences between researchers on the two sides of the lymphocyte divide. BCR repertoire studies will be transformed when greater attention is paid to antigen-specific lineages. TCR repertoire studies, in turn, could benefit from the lessons of the BCR repertoire, which suggest that the analysis of full-length V(D)J rearrangements, and detailed analysis of the nucleotide elements within the CDR3, can help explain the shaping of the repertoire.

ACKNOWLEDGMENTS

This work was made possible by a grant from the National Health and Medical Research Council.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) **302**:575–81. doi:10.1038/302575a0
2. Schroeder HW Jr, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol* (2010) **125**:S41–52. doi:10.1016/j.jaci.2009.09.046
3. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* (1988) **334**:395–402. doi:10.1038/334395a0
4. Takihara Y, Tkachuk D, Michalopoulos E, Champagne E, Reimann J, Minden M, et al. Sequence and organization of the diversity, joining, and constant region genes of the human T-cell delta-chain locus. *Proc Natl Acad Sci U S A* (1988) **85**:6097–101. doi:10.1073/pnas.85.16.6097
5. Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* (2004) **4**:123–32. doi:10.1038/nri1292
6. Desiderio SV, Yancopoulos GD, Paskind M, Thomas E, Boss MA, Landau N, et al. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature* (1984) **311**:752–5.
7. Davies JM, Platts-Mills TA, Aalberse RC. The enigma of IgE+ B-cell memory in human subjects. *J Allergy Clin Immunol* (2013) **131**:972–6. doi:10.1016/j.jaci.2012.12.1569
8. Fukui K, Noma T, Takeuchi K, Kobayashi N, Hatanaka M, Honjo T. Origin of adult T-cell leukemia virus. Implication for its zoonosis. *Mol Biol Med* (1983) **1**:447–56.
9. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun* (2012) **13**:363–73. doi:10.1038/gene.2012.12
10. Rada C, Milstein C. The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. *EMBO J* (2001) **20**:4570–6. doi:10.1093/emboj/20.16.4570
11. Odegard VH, Schatz DG. Targeting of somatic hypermutation. *Nat Rev Immunol* (2006) **6**:573–83. doi:10.1038/nri1896
12. Pallares N, Lefebvre S, Contet V, Matsuda F, Lefranc MP. The human immunoglobulin heavy variable genes. *Exp Clin Immunogenet* (1999) **16**:36–60. doi:10.1159/000019095
13. Ohm-Laursen L, Larsen SR, Barington T. Identification of two new alleles, IGHV3-23*04 and IGHJ6*04, and the complete sequence of the IGHV3-h pseudogene in the human immunoglobulin locus and their prevalences in Danish Caucasians. *Immunogenetics* (2005) **57**:621–7. doi:10.1007/s00251-005-0035-8
14. Romo-Gonzalez T, Morales-Montor J, Rodriguez-Dorantes M, Vargas-Madrazo E. Novel substitution polymorphisms of human immunoglobulin VH genes in Mexicans. *Hum Immunol* (2005) **66**:732–40. doi:10.1016/j.humimm.2005.03.002
15. Wang Y, Jackson KJ, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol* (2008) **86**:111–5.
16. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, et al. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* (2011) **63**:259–65. doi:10.1007/s00251-010-0510-8
17. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) **92**:530–46. doi:10.1016/j.ajhg.2013.03.004
18. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) **184**:6986–92. doi:10.4049/jimmunol.1000445
19. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* (2012) **188**:1333–40. doi:10.4049/jimmunol.1102097
20. Collins AM, Wang Y, Singh V, Yu P, Jackson KJ, Sewell WA. The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. *Immunogenetics* (2008) **60**:669–76. doi:10.1007/s00251-008-0325-z
21. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to *Haemophilus influenzae* type b disease. *J Clin Invest* (1996) **97**:2277–82. doi:10.1172/JCI118669
22. Rowen L, Koop BF, Hood L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* (1996) **272**:1755–62. doi:10.1126/science.272.5269.1755

23. Boysen C, Simon MI, Hood L. Analysis of the 1.1-Mb human alpha/delta T-cell receptor locus with bacterial artificial chromosome clones. *Genome Res* (1997) **7**:330–8.
24. Ibberson MR, Copier JP, Llop E, Navarrete C, Hill AV, Cruickshank JK, et al. T-cell receptor variable alpha (TCRAV) polymorphisms in European, Chinese, South American, AfroCaribbean, and Gambian populations. *Immunogenetics* (1998) **47**:124–30. doi:10.1007/s002510050337
25. Moody AM, Reyburn H, Willcox N, Newsom-Davis J. New polymorphism of the human T-cell receptor AV28S1 gene segment. *Immunogenetics* (1998) **48**:62–4. doi:10.1007/s002510050401
26. Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am J Hum Genet* (2001) **69**:381–95. doi:10.1086/321297
27. Mackelprang R, Livingston RJ, Eberle MA, Carlson CS, Yi Q, Akey JM, et al. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. *Hum Genet* (2006) **119**:255–66. doi:10.1007/s00439-005-0111-z
28. Vessey SJ, Bell JI, Jakobsen BK. A functionally significant allelic polymorphism in a T cell receptor V beta gene segment. *Eur J Immunol* (1996) **26**:1660–3. doi:10.1002/eji.1830260739
29. Gras S, Chen Z, Miles JJ, Liu YC, Bell MJ, Sullivan LC, et al. Allelic polymorphism in the T cell receptor and its impact on immune responses. *J Exp Med* (2010) **207**:1555–67. doi:10.1084/jem.20100603
30. Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J Mol Biol* (1997) **270**:587–97. doi:10.1006/jmbi.1997.1141
31. Munshaw S, Kepler TB. SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* (2010) **26**:867–72. doi:10.1093/bioinformatics/btq056
32. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013). doi:10.1093/nar/gkt382
33. Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* (2007) **23**:1580–7. doi:10.1093/bioinformatics/btm147
34. Jackson KJ, Boyd S, Gaeta BA, Collins AM. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics* (2010) **26**:3129–30. doi:10.1093/bioinformatics/btq604
35. Basu M, Hegde MV, Modak MJ. Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase. *Biochem Biophys Res Commun* (1983) **111**:1105–12. doi:10.1016/0006-291X(83)91413-4
36. Gauss GH, Lieber MR. Mechanistic constraints on diversity in human V(D)J recombination. *Mol Cell Biol* (1996) **16**:258–69.
37. Jackson KJ, Gaeta BA, Collins AM. Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. *J Immunol Methods* (2007) **324**:26–37. doi:10.1016/j.jim.2007.04.011
38. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) **29**:542–50. doi:10.1093/bioinformatics/btt004
39. Lee CE, Jackson KJ, Sewell WA, Collins AM. Use of IGHJ and IGHD gene mutations in analysis of immunoglobulin sequences for the prognosis of chronic lymphocytic leukemia. *Leuk Res* (2007) **31**:1247–52. doi:10.1016/j.leukres.2006.10.013
40. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor Jbeta gene segment usage. *Proc Natl Acad Sci U S A* (2012) **109**:15865–70. doi:10.1073/pnas.1203916109
41. Quiros Roldan E, Sottini A, Bettinardi A, Albertini A, Imberti L, Primi D. Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells. *Immunogenetics* (1995) **41**:91–100.
42. Livak F, Burtrum DB, Rowen L, Schatz DG, Petrie HT. Genetic modulation of T cell receptor gene segment usage during somatic recombination. *J Exp Med* (2000) **192**:1191–6. doi:10.1084/jem.192.8.1191
43. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
44. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* (2003) **17**:2257–317. doi:10.1038/sj.leu.2403202
45. Matthews C, Catherwood M, Morris TC, Alexander HD. Routine analysis of IgVH mutational status in CLL patients using BIOMED-2 standardized primers and protocols. *Leuk Lymphoma* (2004) **45**:1899–904. doi:10.1080/10428190410001710812
46. Glanville J, Kuo TC, Von Büdingen H-C, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci U S A* (2011) **108**:20066–71. doi:10.1073/pnas.1107498108
47. Briney BS, Willis JR, Mckinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes Immun* (2012) **13**:469–73. doi:10.1038/gene.2012.20
48. Benichou J, Glanville J, Prak ET, Azran R, Kuo TC, Pons J, et al. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol* (2013) **190**:5567–77. doi:10.4049/jimmunol.1201929
49. Yamada M, Wasserman R, Reichard BA, Shane S, Caton AJ, Rovera G. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. *J Exp Med* (1991) **173**:395–407. doi:10.1084/jem.173.2.395
50. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res* (2008) **4**:3. doi:10.1186/1745-7580-4-3
51. Klein R, Jaenichen R, Zachau HG. Expressed human immunoglobulin kappa genes and their hypermutation. *Eur J Immunol* (1993) **23**:3248–62. doi:10.1002/eji.1830231231
52. Cox JP, Tomlinson IM, Winter G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. *Eur J Immunol* (1994) **24**:827–36. doi:10.1002/eji.1830240409
53. Weber JC, Blaison G, Martin T, Knapp AM, Pasquali JL. Evidence that the V kappa III gene usage is nonstochastic in both adult and newborn peripheral B cells and that peripheral CD5+ adult B cells are oligoclonal. *J Clin Invest* (1994) **93**:2093–105. doi:10.1172/JCI117204
54. Foster SJ, Brezinschek HP, Brezinschek RI, Lipsky PE. Molecular mechanisms and selective influences that shape the kappa gene repertoire of IgM+ B cells. *J Clin Invest* (1997) **99**:1614–27. doi:10.1172/JCI119324
55. Jackson KJ, Wang Y, Gaeta BA, Pomat W, Siba P, Rimmer J, et al. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* (2012) **64**:3–14. doi:10.1007/s00251-011-0559-z
56. Gay D, Saunders T, Camper S, Weigert M. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J Exp Med* (1993) **177**:999–1008. doi:10.1084/jem.177.4.999
57. Tiegs SL, Russell DM, Nemazee D. Receptor editing in self-reactive bone marrow B cells. *J Exp Med* (1993) **177**:1009–20. doi:10.1084/jem.177.4.1009
58. Mehr R, Shannon M, Litwin S. Models for antigen receptor gene rearrangement. I. Biased receptor editing in B cells: implications for allelic exclusion. *J Immunol* (1999) **163**:1793–8.
59. Ignatovich O, Tomlinson IM, Jones PT, Winter G. The creation of diversity in the human immunoglobulin V(lambda) repertoire. *J*

Mol Biol (1997) **268**:69–77. doi:10.1006/jmbi.1997.0956
60. Vasicek TJ, Leder P. Structure and expression of the human immunoglobulin lambda genes. *J Exp Med* (1990) **172**:609–20. doi:10.1084/jem.172.2.609
61. Farner NL, Dorner T, Lipsky PE. Molecular mechanisms and selection influence the generation of the human V lambda J lambda repertoire. *J Immunol* (1999) **162**:2137–45.
62. Tuaillon N, Taylor LD, Lonberg N, Tucker PW, Capra JD. Human immunoglobulin heavy-chain minilocus recombination in transgenic mice: gene-segment use in mu and gamma transcripts *Proc Natl Acad Sci U S A* (1993) **90**:3720–4. doi:10.1073/pnas.90.8.3720
63. Hoi KH, Ippolito GC. Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. *Genes Immun* (2013) **14**:271–6. doi:10.1038/gene.2013.10
64. Craddock TP, Zumla AM, Ollier WE, Chintu CZ, Muyinda GP, Lancaster FC, et al. Predominance of one T-cell antigen receptor BV haplotype in African populations. *Immunogenetics* (2000) **51**:231–7. doi:10.1007/s002510050036
65. Donaldson IJ, Shefta J, Lawson CA, Bushnell JR, Morgan AW, Isaacs JD, et al. Unique TCR beta-subunit variable gene haplotypes in Africans. *Immunogenetics* (2002) **53**:884–93. doi:10.1007/s00251-001-0406-8
66. McMurry MT, Hernandez-Munain C, Lauzurica P, Krangel MS. Enhancer control of local accessibility to V(D)J recombinase. *Mol Cell Biol* (1997) **17**:4553–61.
67. Nadel B, Tang A, Escuro G, Lugo G, Feeney AJ. Sequence of the spacer in the recombination signal sequence affects V(D)J rearrangement frequency and correlates with nonrandom Vkappa usage in vivo. *J Exp Med* (1998) **187**:1495–503. doi:10.1084/jem.187.9.1495
68. Feeney AJ, Tang A, Ogwaro KM. B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. *Immunol Rev* (2000) **175**:59–69. doi:10.1111/j.1600-065X.2000.imr017508.x
69. Posnett DN, Vissinga CS, Pambuccian C, Wei S, Robinson MA, Kostyu D, et al. Level of human TCRBV3S1 (V beta 3) expression correlates with allelic polymorphism in the spacer region of the recombination signal sequence. *J Exp Med* (1994) **179**:1707–11. doi:10.1084/jem.179.5.1707
70. Wei Z, Lieber MR. Lymphoid V(D)J recombination. Functional analysis of the spacer sequence within the recombination signal. *J Biol Chem* (1993) **268**:3180–3.
71. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* (1998) **188**:2151–62. doi:10.1084/jem.188.11.2151
72. Sasso EH, Willems Van Dijk K, Bull A, Van Der Maarel SM, Milner EC. VH genes in tandem array comprise a repeated germline motif. *J Immunol* (1992) **149**:1230–6.
73. Sasso EH, Johnson T, Kipps TJ. Expression of the immunoglobulin VH gene 51p1 is proportional to its germline gene copy number. *J Clin Invest* (1996) **97**:2074–80. doi:10.1172/JCI118644
74. Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol* (2004) **172**:6790–802.
75. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* (2009) **1**:12ra23. doi:10.1126/scitranslmed.3000540
76. De Preval C, Fougereau M. Specific interaction between VH and VL regions of human monoclonal immunoglobulins. *J Mol Biol* (1976) **102**:657–78. doi:10.1016/0022-2836(76)90340-5
77. Brezinschek HP, Foster SJ, Dorner T, Brezinschek RI, Lipsky PE. Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J Immunol* (1998) **160**:4762–7.
78. de Wildt RM, Hoet RM, Van Venrooij WJ, Tomlinson IM, Winter G. Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. *J Mol Biol* (1999) **285**:895–901. doi:10.1006/jmbi.1998.2396
79. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) **31**:166–9. doi:10.1038/nbt.2492
80. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE* (2011) **6**:e22365. doi:10.1371/journal.pone.0022365
81. Roth DB, Menetski JP, Nakajima PB, Bosma MJ, Gellert M. V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. *Cell* (1992) **70**:983–91. doi:10.1016/0092-8674(92)90248-B
82. Lafaille JJ, Decloux A, Bonneville M, Takagaki Y, Tonegawa S. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell* (1989) **59**:859–70. doi:10.1016/0092-8674(89)90609-0
83. Tian C, Luskin GK, Dischert KM, Higginbotham JN, Shepherd BE, Crowe JE Jr. Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naive and memory B cells. *Mol Immunol* (2007) **44**:2173–83. doi:10.1016/j.molimm.2006.11.020
84. Jackson KJ, Gaeta B, Sewell W, Collins AM. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol* (2004) **5**:19. doi:10.1186/1471-2172-5-19
85. Lu H, Schwarz K, Lieber MR. Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res* (2007) **35**:6917–23. doi:10.1093/nar/gkm823
86. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* (2001) **194**:1385–90. doi:10.1084/jem.194.9.1385
87. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* (2012) **109**:16161–6. doi:10.1073/pnas.1212755109
88. Hofle M, Linthicum DS, Ioerger T. Analysis of diversity of nucleotide and amino acid distributions in the VD and DJ joining regions in Ig heavy chains. *Mol Immunol* (2000) **37**:827–35. doi:10.1016/S0161-5890(00)00110-3
89. Boubnov NV, Wills ZP, Weaver DT. V(D)J recombination coding junction formation without DNA homology: processing of coding termini. *Mol Cell Biol* (1993) **13**:6957–68.
90. Nadel B, Feeney AJ. Influence of coding-end sequence on coding-end processing in V(D)J recombination. *J Immunol* (1995) **155**:4322–9.
91. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. *Immunol Cell Biol* (2007) **85**:323–32. doi:10.1038/sj.icb.7100055
92. Agathangelidis A, Darzentas N, Hadzidimitriou A, Brochet X, Murray F, Yan XJ, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood* (2012) **119**:4467–75. doi:10.1182/blood-2011-11-393694
93. Hoogeboom R, Van Kessel KP, Hochstenbach F, Wormhoudt TA, Reinten RJ, Wagner K, et al. A mutated B cell chronic lymphocytic leukemia subset that recognizes and responds to fungi. *J Exp Med* (2013) **210**:59–70. doi:10.1084/jem.20121801
94. Smith K, Garman L, Wrammert J, Zheng NY, Capra JD, Ahmed R, et al. Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat Protoc* (2009) **4**:372–84. doi:10.1038/nprot.2009.3
95. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* (2013) **5**:171ra119. doi:10.1126/scitranslmed.3004794
96. Wu YC, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front Immunol* (2012) **3**:193. doi:10.3389/fimmu.2012.00193
97. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) **333**:1593–602. doi:10.1126/science.1207532

98. Parameswaran P, Yi Liu Q, Roskin KM, Jackson KK, Dixit VP, Lee J-Y, et al. Coherent immune repertoire signatures in human dengue. *Cell Host Microbe* (Forthcoming 2013).
99. Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, Bodmer HC, Mcmichael AJ, et al. Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. *Proc Natl Acad Sci U S A* (1991) **88**:8987–90. doi:10.1073/pnas.88.20.8987
100. Argaet VP, Schmidt CW, Burrows SR, Silins SL, Kurilla MG, Doolan DL, et al. Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus. *J Exp Med* (1994) **180**:2335–40. doi:10.1084/jem.180.6.2335
101. Pannetier C, Even J, Kourilsky P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol Today* (1995) **16**:176–81. doi:10.1016/0167-5699(95)80117-0
102. Altman JD, Moss PA, Goulder PJ, Barouch DH, Mcheyzer-Williams MG, Bell JI, et al. Phenotypic analysis of antigen-specific T lymphocytes. *Science* (1996) **274**:94–6. doi:10.1126/science.274.5284.94
103. Garrido P, Ruiz-Cabello F, Barcena P, Sandberg Y, Canton J, Lima M, et al. Monoclonal TCR-Vbeta13.1+/CD4+/NKa+/CD8-/+dim T-LGL lymphocytosis: evidence for an antigen-driven chronic T-cell stimulation origin. *Blood* (2007) **109**:4890–8. doi:10.1182/blood-2006-05-022277
104. Crompton L, Khan N, Khanna R, Nayak L, Moss PA. CD4+ T cells specific for glycoprotein B from cytomegalovirus exhibit extreme conservation of T-cell receptor usage between different individuals. *Blood* (2008) **111**:2053–61. doi:10.1182/blood-2007-04-079863
105. Moon JJ, Chu HH, Pepper M, Mcsorley SJ, Jameson SC, Kedl RM, et al. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* (2007) **27**:203–13. doi:10.1016/j.immuni.2007.07.007
106. Alanio C, Lemaitre F, Law HK, Hasan M, Albert ML. Enumeration of human antigen-specific naive CD8+ T cells reveals conserved precursor frequencies. *Blood* (2010) **115**:3718–25. doi:10.1182/blood-2009-10-251124
107. Kwok WW, Tan V, Gillette L, Littell CT, Soltis MA, Lafond RB, et al. Frequency of epitope-specific naive CD4(+) T cells correlates with immunodominance in the human memory repertoire. *J Immunol* (2012) **188**:2537–44. doi:10.4049/jimmunol.1102190
108. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci USA* (2006) **103**:18691–6. doi:10.1073/pnas.0608907103
109. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci USA* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107
110. Miles JJ, Douek DC, Price DA. Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* (2011) **89**:375–87. doi:10.1038/icb.2010.139

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2013; accepted: 19 August 2013; published online: 02 September 2013.

*Citation: Jackson KJL, Kidd MJ, Wang Y and Collins AM (2013) The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front. Immunol. **4**:263. doi: 10.3389/fimmu.2013.00263*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Jackson, Kidd, Wang and Collins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 November 2013
doi: 10.3389/fimmu.2013.00358

Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data

Gur Yaari[1,2], Jason A. Vander Heiden[3], Mohamed Uduman[2], Daniel Gadala-Maria[3], Namita Gupta[3], Joel N. H. Stern[4,5], Kevin C. O'Connor[4,6], David A. Hafler[4,7], Uri Laserson[8], Francois Vigneault[9] and Steven H. Kleinstein[2,3]*

[1] *Bioengineering Program, Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel*
[2] *Department of Pathology, Yale School of Medicine, New Haven, CT, USA*
[3] *Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA*
[4] *Department of Neurology, Yale School of Medicine, New Haven, CT, USA*
[5] *Department of Science Education, Hofstra North Shore-LIJ School of Medicine, Hempstead, NY, USA*
[6] *Human and Translational Immunology Program, Yale School of Medicine, New Haven, CT, USA*
[7] *Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA*
[8] *Department of Genetics, Harvard Medical School, Boston, MA, USA*
[9] *AbVitro, Inc., Boston, MA, USA*

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ronald B. Corley, Boston University School of Medicine, USA
Masaki Hikida, Kyoto University, Japan

***Correspondence:**
Steven H. Kleinstein, Department of Pathology, Yale School of Medicine, Suite 505, 300 George Street, New Haven, CT 06511, USA
e-mail: steven.kleinstein@yale.edu

Analyses of somatic hypermutation (SHM) patterns in B cell immunoglobulin (Ig) sequences contribute to our basic understanding of adaptive immunity, and have broad applications not only for understanding the immune response to pathogens, but also to determining the role of SHM in autoimmunity and B cell cancers. Although stochastic, SHM displays intrinsic biases that can confound statistical analysis, especially when combined with the particular codon usage and base composition in Ig sequences. Analysis of B cell clonal expansion, diversification, and selection processes thus critically depends on an accurate background model for SHM micro-sequence targeting (i.e., hot/cold-spots) and nucleotide substitution. Existing models are based on small numbers of sequences/mutations, in part because they depend on data from non-coding regions or non-functional sequences to remove the confounding influences of selection. Here, we combine high-throughput Ig sequencing with new computational analysis methods to produce improved models of SHM targeting and substitution that are based only on synonymous mutations, and are thus independent of selection. The resulting "S5F" models are based on 806,860 Synonymous mutations in 5-mer motifs from 1,145,182 Functional sequences and account for dependencies on the adjacent four nucleotides (two bases upstream and downstream of the mutation). The estimated profiles can explain almost half of the variance in observed mutation patterns, and clearly show that both mutation targeting and substitution are significantly influenced by neighboring bases. While mutability and substitution profiles were highly conserved across individuals, the variability across motifs was found to be much larger than previously estimated. The model and method source code are made available at http://clip.med.yale.edu/SHM

Keywords: immunoglobulin, B cell, somatic hypermutation, mutability, substitution, targeting, AID, affinity maturation

1. INTRODUCTION

During the course of an immune response, B cells that initially bind antigen with low affinity through their immunoglobulin (Ig) receptor are modified through cycles of proliferation, somatic hypermutation (SHM), and affinity-dependent selection to produce high-affinity memory and plasma cells. Current models of SHM recognize activation-induced deaminase (AID), along with several DNA repair pathways, as critical to the mutation process (1). AID initiates SHM by converting cytosines (Cs) to uracils (Us), thus creating U:G mismatches in the Ig V(D)J sequence. If not repaired before cell replication, these mismatches produce C → T (thymine) transition mutations (2). The AID-induced mismatches can alternatively be recognized by UNG or MSH2/MSH6 to initiate base excision or mismatch repair pathways, respectively. These pathways operate in an error-prone manner to introduce the full spectrum of mutations at the initial lesion, as well as spreading mutations to the surrounding bases. Overall, SHM introduces point mutations into the Ig locus at a rate of $\sim 10^{-3}$ per base-pair per division (3, 4). While the process of SHM appears to be stochastic, there are clear intrinsic biases, both in the bases that are targeted (5, 6) as well as the substitutions that are introduced (7, 8). Accurate background models for SHM micro-sequence targeting

(i.e., hot/cold-spots) and nucleotide substitution would greatly aid the analysis of B cell clonal expansion, diversification, and selection processes. In addition, targeting and substitution models could provide important insights into the relative contributions of the various error-prone DNA repair pathways that mediate SHM.

Computational models and analyses of SHM have separated the process into two independent components (7–11): (1) a targeting model that defines where mutations occur (by specifying the relative rates at which positions in the Ig sequence are mutated), and (2) a nucleotide substitution model that defines the resulting mutation (by specifying the probability of each base mutating to each of the other three possibilities). In experimentally derived Ig sequences, observed mutation patterns are influenced by selection. The affinity maturation process selects for affinity-increasing mutations, while many mutations at structurally important positions in the framework regions are selected against (12). To avoid the confounding influences of selection, most existing models are built using mutation data from intronic regions flanking the V gene (13) and non-productively rearranged Ig genes (6–10, 14). These works have identified several specific motifs as being "hot" or "cold" spots of SHM. Hot-spots include WRCY/RGYW and WA/TW (where W = {A, T}, Y = {C, T} R = {G, A}, and the mutated position is underlined, see for example (5, 6)). Although it has been argued that WRCH/DGYW (where H = {A, C, T} and D = {A, G, T}) is a better predictor of mutability at C:G bases (15). A single cold-spot motif has also been recognized: SYC/GRS (where S = {C, G}) (16). Despite the wide recognition of these specific hot-spot and cold-spot motifs, it is clear that a hierarchy of mutabilities exists that is highly dependent on the surrounding bases (7, 10). More recently, it has been recognized that the profile of nucleotide substitutions may also be dependent on the surrounding bases (8, 17). Modeling SHM targeting and substitution is important for the analysis of mutation patterns since these intrinsic biases can give the appearance of selection due to the particular codon usage and base composition in Ig sequences (17, 18). Moreover, having such a model could shed light on the molecular mechanisms underlying SHM, which are not fully understood.

Previous work has attempted to model the dependencies on surrounding bases, but has been limited to (at most) the targeted base and three surrounding bases (19), mainly due to the relatively small data sets available. The use of intronic regions has also limited the number of motifs that can be modeled (because of the limited diversity of these regions), and non-productively rearranged Ig genes may still be influenced by selection (e.g., if the event rendering the sequence non-productive happened in the course of affinity maturation). In this study, we take advantage of the wealth of data available from high-throughput Ig sequencing technologies to build improved targeting and substitution models for SHM. To avoid the biasing effects of selection, we have developed a new methodology for constructing models from synonymous mutations only, thus avoiding the need to limit analysis to non-productive Ig sequences. The increased data set size allows modeling of dependencies on the surrounding four bases (two bases upstream and downstream of the mutation). These "S5F" (Synonymous, 5-mer, Functional) models confirm the existence of proposed hot- and cold-spots of SHM, but also show much more extreme difference between hot- and cold-spots compared with previous models. We also find that the nucleotide substitution profiles at all bases are dependent on the surrounding nucleotides. The S5F targeting and substitution models can be employed as background distributions for mutation analysis, such as the detection and quantification of affinity-dependent selection in Ig sequences (11, 20). These models improve dramatically the ability to analyze mutation patterns in Ig sequences, and provide insights into the SHM process.

2. RESULTS

To develop models for SHM targeting and substitution preferences, we curated a large database of mutations from high-throughput sequencing studies (**Table 1**). These data were derived from 7 human blood and lymph node samples, and Ig sequencing was carried out using both Roche 454 and Illumina MiSeq next-generation sequencing technologies. In total, the data contained 42,122,509 raw reads, which were processed (see Materials and Methods) to arrive at 1,145,182 "high-fidelity" Ig sequences, which were each supported by a minimum of two independent reads in a sample. These high-fidelity sequences were clustered to identify clones (sequences related by a common ancestor) and one effective sequence was constructed per clone so that each observed mutation corresponded to an independent event. Overall, this process produced a set of 806,860 synonymous mutations that were used to model somatic hypermutation targeting and substitution.

2.1. THE NUCLEOTIDE SUBSTITUTION SPECTRUM IS AFFECTED BY ADJACENT NUCLEOTIDES

A nucleotide substitution model specifies the probability of each base (A, T, G, or C) mutating to each of the other three possibilities. For example, when a C is mutated, we might find that 50% of the time it is replaced by T, while 30% of the substitutions are to G, and the remaining 20% lead to A. These probabilities may depend on the surrounding bases (i.e., the micro-sequence context), as was previously suggested for mutations at A (17) and more generally (8). To derive a nucleotide substitution model, the set of mutations was filtered to include only those that occurred in positions where none of the possible base substitutions lead to amino acid exchanges. Focusing on positions where only synonymous mutations were possible removes the confounding influence of selection. The resulting 408,422 mutations were analyzed and grouped into "5-mers" according to the germline sequence of the mutated position and surrounding bases (two base-pairs upstream and two base-pairs downstream of the mutated position). For each of the 1024 possible 5-mers (M), a substitution model was derived by calculating S_B^M, the probability that the central base in the 5-mer motif (M) mutates to base B. For example, in the 5-mer CCATC mutations at A are always synonymous whenever this motif starts a reading frame, in which case it codes for a Proline (CCA) followed by a Serine (TCN). In this case, the number of observed mutations that led to each of the other three possible nucleotides (C, G, or T) was recorded: N_C^{CCATC}, N_G^{CCATC}, N_T^{CCATC}. The maximum likelihood value for the probability that A is substituted by base B is then calculated as:

$$S_B^{CCATC} = \frac{N_B^{CCATC}}{N_C^{CCATC} + N_G^{CCATC} + N_T^{CCATC}}$$

Table 1 | Next-generation sequencing data sets used to construct the "S5F" targeting and substitution models.

Study	Sample	Subject	Tissue	Tech.	Raw reads	Processed reads	Clones	# Mutations (substitution)	# Mutations (targeting)
1	3931LN	1	LN	MiSeq	3,641,633	79,777	16,272	25,307	53,840
1	4014LN	2	LN	MiSeq	3,714,152	106,006	32,972	57,215	106,265
1	4106LN	3	LN	MiSeq	10,917,517	231,387	54,400	108,591	208,338
1	3928LN	4	LN	MiSeq	7,691,509	99,519	76,375	68,051	132,795
2	PGP1-1	5	PBMC	MiSeq	3,851,658	55,606	50,514	23,939	48,558
2	PGP1-2	5	PBMC	MiSeq	3,946,514	59,611	54,374	24,971	50,117
2	PGP1-3	5	PBMC	MiSeq	4,543,353	48,971	45,788	20,865	42,737
2	PGP1-4	5	PBMC	MiSeq	3,121,884	52,844	49,054	23,243	47,049
3	hu420143	6	PBMC	454	178,584	92,055	14,956	23,260	48,838
3	420IV	7	PBMC	454	398,517	248,363	39,047	24,771	50,899
3	PGP1-5	5	PBMC	454	117,188	71,043	12,275	8,209	17,424
Total	–	–	–	–	42,122,509	1,145,182	446,027	408,422	806,860

Tissue types are lymph node (LN) or peripheral blood mononuclear cell (PBMC). The different filters applied to arrive at the number of (synonymous) mutations used for the targeting and substitution models are described in the text. All three studies relate to manuscripts in preparation.

A bootstrapping procedure was used to estimate 95% confidence intervals (21).

Comparison of the substitution profiles for different 5-mer motifs with the same central base clearly showed the significant influence of surrounding bases. As an example, **Figure 1A** shows how the profile of substitutions at G changes for several different 5-mers (AC<u>G</u>AT, GC<u>G</u>AG, GT<u>G</u>TA, and GG<u>G</u>AA). Such dependencies were identified for every base (A, T, G, and C) (**Figure 1B** and Figure S1 in Supplementary Material). The importance of including two bases upstream and downstream was confirmed by comparing these profiles with analogous profiles that only account for the immediately adjacent bases (3-mer motifs) (**Figure 1A**). For the 3-mer C<u>G</u>A, G → C and G → A substitutions were equally likely (45% and 43% of substitutions, respectively), while G → C substitutions were significantly more likely than G → A in the context of the GC<u>G</u>AG motif (51% and 35% of substitutions, respectively). If one ignores neighboring nucleotides, the substitution profiles were qualitatively similar to previous estimates (7), although significant quantitative differences were apparent (presumably due to the much larger size of the dataset compiled here). Thus, nucleotide substitution profiles at every base are significantly affected by adjacent nucleotides, including at least two bases on either side of the mutating base.

2.1.1. The complete substitution model for somatic hypermutation is not strand-symmetric

It is not possible to estimate substitution profiles for all 5-mer motifs using the above methodology because: (1) not all 5-mers appear within the set of Ig sequences, and (2) some 5-mers (such as NA<u>N</u>NN) can never appear in a context where all substitutions at the central (underlined) base are synonymous. Among the 11 datasets used here, these issues prevent estimation of the substitution profiles for 717 of the 1024 5-mers. For the profiles that could be directly estimated, there was a high correlation (on average Pearson $R = 0.63$) between different individuals (Figures S2 and S3 in Supplementary Material), and so all the samples were combined to estimate a single substitution model. To infer values for the missing motifs, four methods were evaluated. In the first method ("inner 3-mer"), the substitution profile for each missing 5-mer was inferred by averaging over profiles for all 5-mers with the same 3-mer core (i.e., for which the middle three bases were shared). In the second and third methods, missing values were replaced by averaging over motifs sharing the two bases upstream and downstream of the mutated base, respectively. In the fourth method ("hot-spot"), the missing substitution profile was inferred by averaging over 5-mers sharing the two upstream bases when the mutated position was "C" or "A," and two downstream bases when the mutated position was "G" or "T." This final option was motivated by the dependencies of known "hot" and "cold" spots for SHM targeting (5, 6). To choose between these four methods, we compared their performance on 5-mers that could be directly estimated from the data. Specifically, we calculated the correlation between the inferred and directly estimated ratios for the parameter *R*, which was defined as the ratio between the highest substitution probability with the next highest one for a given 5-mer (**Table 2**). Pearson and Spearman coefficients were both used in order to be robust to the linear dependency assumption, and they yielded comparable results. While the "hot-spot" method clearly had the worst performance, the other three methods resulted in very similar models. The "inner 3-mer" method produced the highest Pearson correlation (0.4, see **Table 2**) and was chosen as the basis to infer missing values. We refer to the resulting substitution model as a "S5F" model since it is based on Synonymous mutations at 5-mers in Functional Ig sequences. In contrast to previous studies (8), there was no significant correlation between substitution values of 5-mers and their reverse complements (Pearson correlation of 0.005, Spearman correlation of 0.087), suggesting that at least one component of the substitution mechanism is not strand-symmetric.

2.2. THE HIERARCHY OF MOTIF MUTABILITIES IS CONSERVED ACROSS INDIVIDUALS

The mutability of a motif is defined here as the (non-normalized) probability of the central base in the motif being targeted for SHM relative to all other motifs. Similar to the substitution model, the targeting model was based on 5-mer motifs, including the

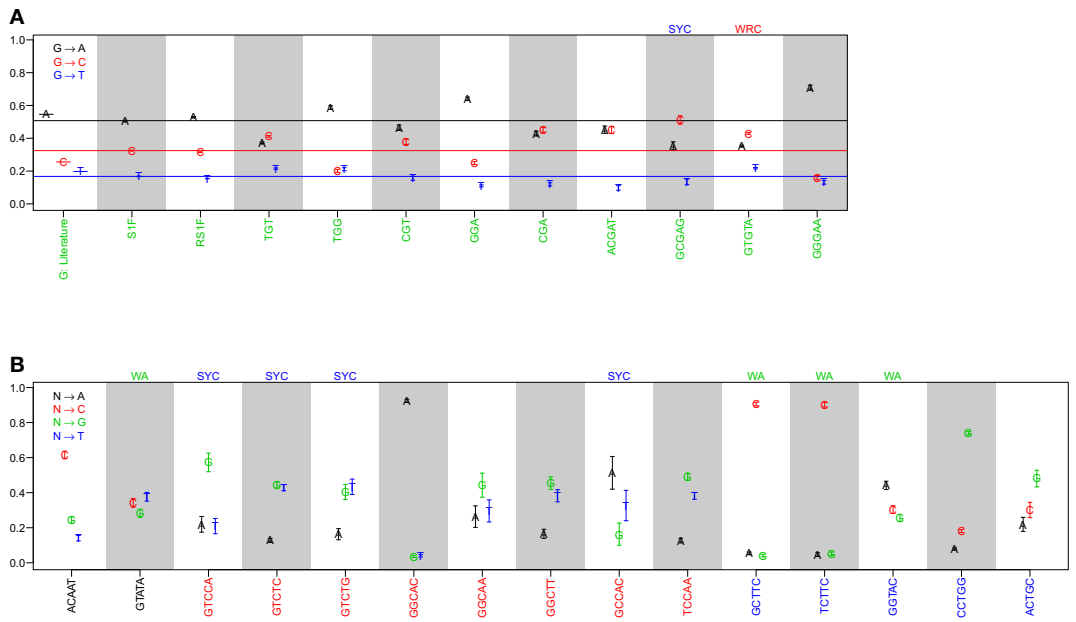


FIGURE 1 | The substitution profile is significantly influenced by surrounding bases. Substitution profiles for various micro-sequence contexts are shown for substitutions at **(A)** guanine and **(B)** adenosine, cytidine, and thymidine. G: literature indicates values estimated by Smith et al. (7), while S1F and RS1F refer to models estimated using the methods proposed here using all (replacement and silent) or only silent mutations, respectively, and averaging over surrounding bases. 3-mer motifs were estimated using silent mutations and dependencies on the immediately adjacent bases (S3F), while 5-mer motifs refer to the complete S5F model. Horizontal lines in **(A)** indicate the substitution values for the S1F model following the color scheme shown in the legend. Motifs that fall into one of the standard hot or cold-spots categories are indicated by the motif above the column.

Table 2 | Correlation coefficients for inferring missing mutability/substitution values.

Model	Correlation	Middle	Upstream	Downstream	Hot spots
Substitution	Pearson	0.40	0.37	0.15	0.04
	Spearman	0.20	0.24	0.23	0.09
Mutability	Pearson	0.58	0.57	0.61	0.73
	Spearman	0.61	0.58	0.64	0.79

two nucleotides immediately upstream and downstream of the mutated base. The use of a 5-mer model is motivated by the well-known WRCY hot-spot (where the underlined C is targeted for mutation), and its reverse-complement (RGYW) which, when taken together, create dependencies with the two bases on either side of the mutating base.

When estimating the mutability (μ) for a motif (M), it is critical to account for the background frequency of M. To see why this is the case, consider the extreme example of a sequence composed of all C nucleotides. Since all mutations will occur at CCCCC motifs, one might consider this motif a hot-spot, except that its background frequency is 100% so it is actually targeted at the expected frequency. When calculating mutabilities it is also important to avoid statistical artifacts due to heterogeneity (e.g., the Simpson paradox (22)). Thus, Ig sequences were first analyzed individually since each has a different background 5-mer distribution. These individual-sequence targeting models were then combined into a single aggregated targeting model for each data set. Estimating the relative mutabilities of 5-mer motifs for an individual Ig sequence involves two steps: (1) Calculating the background frequency of the different 5-mers based on the germline (unmutated) version of the sequence, and (2) creating a table of the 5-mers that were mutated in the sequence. To avoid the confounding influence of selection, only mutations that were synonymous (i.e., that do not produce an amino acid exchange in the germline context) were included in the analysis. Note that these criteria are slightly different from those used in the substitution model. In the substitution model, mutations were used only where all possible mutations at that position had to be synonymous, while all synonymous mutations were considered for mutabilities (see **Table 1**).

For each of the 1024 possible 5-mers motifs (M) in each Ig sequence, the background frequency (B_{M}) was calculated as follows:

$$B_{\mathrm{M}} = \sum_{i} \sum_{b} S_{b}^{\mathrm{M}} I_{\overrightarrow{GL}}(i, \mathrm{M}, b) \qquad (1)$$

where i is summed over all (non-N) positions in the Ig sequence, M is the 5-mer nucleotide sequence centered at position i and b includes all possible nucleotides ({A, C, T, G}). In this equation GL is a vector containing the nucleic content of each position in the germline sequence, S_{b}^{M} is the relative rate at which the center nucleotide in M ($GL[i]$) mutates to b (as estimated in the previous section, and where $S_{GL[i]}^{\mathrm{M}} = 0$) and $I_{\overrightarrow{GL}}(i, M, b)$ is an indicator function that is 1 in cases where the 5-mer surrounding $GL[i]$

is M and a mutation in position i from $GL[i]$ to b results in a synonymous mutation (and 0 otherwise). A similar array was also calculated for the mutated positions:

$$C_{\mathrm{M}} = \sum_{i} I_{\overrightarrow{GL},\overrightarrow{OS}}(i, \mathrm{M}) \quad (2)$$

where i is summed over all (non-N) positions in the observed Ig sequence (OS), and the indicator function $I_{\overrightarrow{GL},\overrightarrow{OS}}$ (i, M) is 1 in cases where the 5-mer surrounding $GL[i]$ is M and a mutation in position i from $GL[i]$ to $OS[i]$ is synonymous and 0 otherwise. After calculating the arrays $\vec{C}$ and $\vec{B}$, a mutability score, μ, was defined for each motif M in the vector (for sequence j) as:

$$\mu^{j}_{\mathrm{M}} = C^{j}_{\mathrm{M}} / B^{j}_{\mathrm{M}} \quad (3)$$

which was then normalized to one:

$$\bar{\mu}^{j}_{\mathrm{M}} = \mu^{j}_{\mathrm{M}} / \sum_{\mathrm{m}} \mu^{j}_{\mathrm{m}} \quad (4)$$

where m is an index spanning all positions in $\vec{\mu}^{j}$. Note that μ^{j}_{M} is not defined wherever $B^{j}_{\mathrm{M}} = 0$ (i.e., the motif M does not appear in the Ig sequence, or can not admit any synonymous mutations). Finally, a single mutability score is generated for each 5-mer motif (M) as the weighted average of the mutabilities scores for each sequence j ($\bar{\mu}^{j}_{\mathrm{M}}$), where weights correspond to the number of synonymous mutations in the sequence ($\sum_{\mathrm{M}} C^{j}_{\mathrm{M}}$). This process resulted in an array of (relative) mutabilities, μ_{M} for each of the 5-mers observed in the dataset. The resulting vector was renormalized so that the mean mutability was one.

2.2.1. Inference of missing values to complete targeting model

It was not possible to estimate mutabilities for 468 of the 1024 possible 5-mer motifs because not all 5-mers appeared within the set of Ig sequences. The same four methods tested for inferring missing values in the substitution model were also tested to infer these mutabilities (see 2.1.1 and **Table 2**). The "inner 3-mer" method produced a Pearson correlation of 0.58 (0.61 for Spearman), while the "hot-spot" method had a correlation of 0.73 (0.79 for Spearman). Thus, in contrast to the nucleotide substitution model, mutabilities were best predicted by averaging over 5-mers which shared the two upstream bases when the mutated position was "C" or "A," and two downstream bases when the mutated position was "G" or "T." This result is consistent with the expected influence of the classic SHM hot-spot (WRCY/RGYW).

2.2.2. Targeting is conserved across individuals

To test whether the micro-sequence specificity of SHM was conserved across individuals, separate targeting models were constructed for each of the 11 samples in our study (**Table 1**). Comparison of the motif mutabilities between pairs of samples showed that the models were highly consistent, with Pearson correlation ∼0.9 (**Figure 2** and Figure S4 in Supplementary Material). Thus, we combined the data from all of the samples and generated a single targeting model, with confidence intervals based on the middle 50% quantiles of the mutability across samples. As with the substitution model, we refer to this targeting model as a "S5F" model. In order to visualize this model, we created "hedgehog" plots to display the directly estimated mutability values and the complete S5F model (**Figures 3A,B**, respectively).

2.2.3. The true "hotness" of SHM hot-spots

Visual inspection of the "hedgehog" plots (**Figure 3B**) shows clearly that the S5F model is consistent with known micro-sequence preferences for SHM (5, 6). WRC/GYW and WA/TW hot-spot motifs are generally more mutable, while SYC/GRS cold-spot motifs generally show the lowest mutability. However, the mutability of "hot-spot" motifs was observed to be highly variable. There is a 62.7-fold difference between the most mutable (GGGCA, mutability = 9.56) and least mutable (TGCGA, mutability = 0.15) WRC/GYW hot-spot motif. Indeed, ∼10% of so-called "hot-spots" had mutabilities that were lower than the mean mutability for "neutral" motifs (**Figure 4A**). This high variance was especially obvious when looking at the subset of WRCA/TGYW hot-spot motifs, and may help explain why WRCH/DGYW has been proposed to be a better predictor of mutation at C:G compared with WRCY/RGYW (15). The mutabilities estimated by the S5F approach paint a qualitatively different picture of SHM when compared with those estimated by the existing tri-nucleotide model of Shapiro et al. (10). In the S5F model, the average mutability of motifs that correspond to the WRC/GYW SHM hot-spot was 3.2-fold higher than neutral motifs, and 9.6-fold higher than the mutability of motifs corresponding to the cold-spot SYC/GRS (**Figure 4A**). Using the tri-nucleotide model, hot-spots were only 1.3-fold and 1.6-fold more mutable than neutral and cold-spots, respectively (**Figure 4B**). In addition, in direct opposition to the S5F model, the tri-nucleotide method predicted that A/T hot-spots (WA/TW) were more mutable than C/G hot-spots (WRC/GYW). The mutabilities estimated by the S5F model better predicted the positional-distribution of *in vivo* mutations. The Pearson correlation between the expected mutability and observed mutation frequency calculated over IMGT-numbered positions in 12,000 sequences derived from a variety of germline segments was 0.67 and 0.47 for the S5F and tri-nucleotide models, respectively (**Figure 5** and Figure S5 in Supplementary Material). In both methods, deviations from the expected frequencies that likely reflect both positive and negative selection were observed (**Figure 5**). The observation of position-specific signals suggests that there is something generic about the Ig structure at these positions, and may help refine traditional definitions of the complementarity determining regions (CDR) and framework regions (FWR) (see also (23)). Consistent with previous studies (24), the S5F model displayed significant strand-bias at A/T hot-spots, but not C/G hot-spots (**Figure 6**). Overall, the S5F targeting model provides a new view of SHM with hot-spots being significantly more targeted (and significantly more variable) than previously thought.

3. MATERIALS AND METHODS

3.1. HIGH-THROUGHPUT IG SEQUENCING DATA SETS

A total of 11 human Ig repertoires were sequenced from blood and lymph node samples from 7 different individuals. Next-generation sequencing was carried out using Illumina MiSeq 250

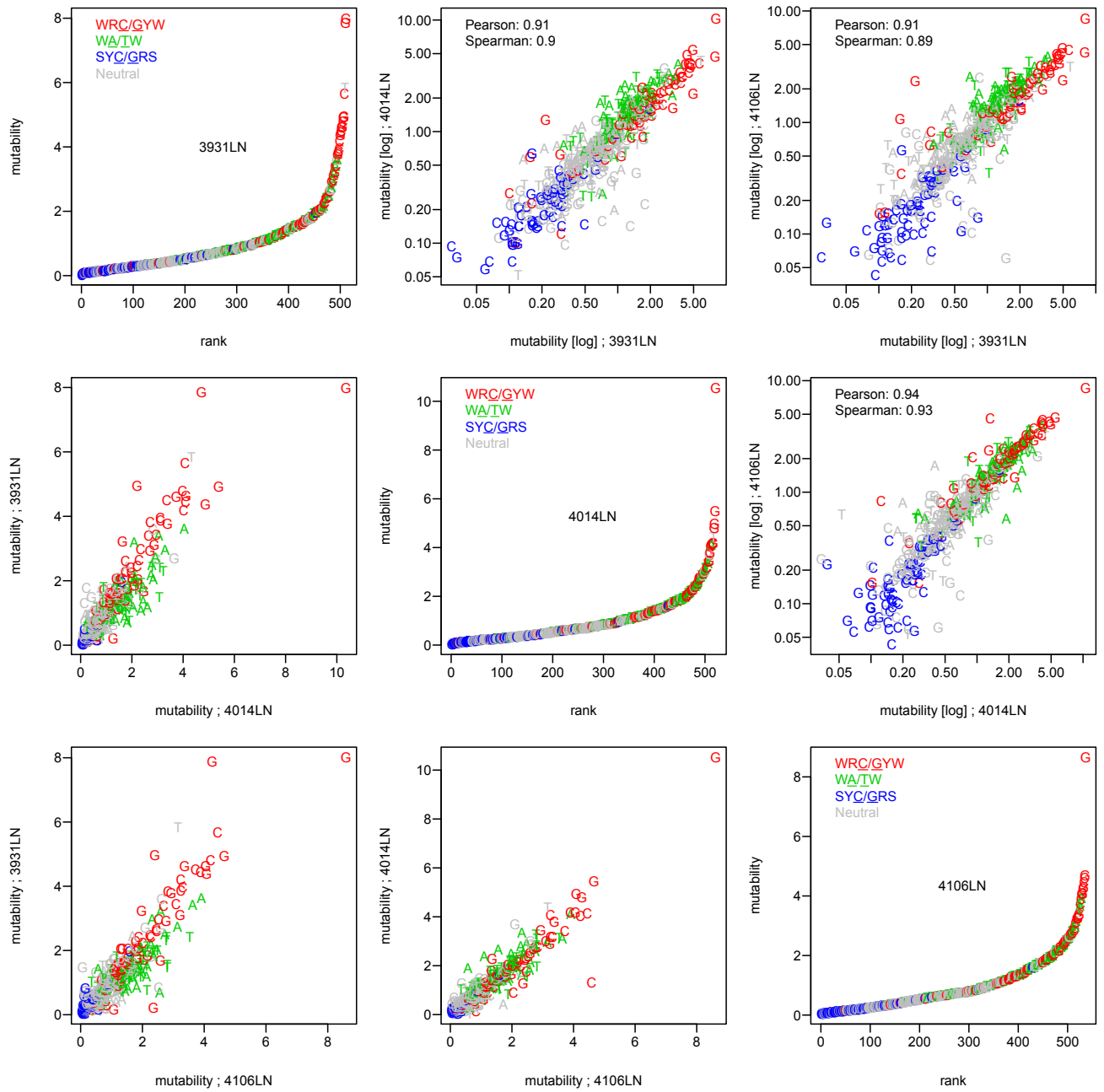


FIGURE 2 | The S5F targeting model is consistent across individuals. Targeting models were constructed independently for each of the samples listed in **Table 1**. Estimated values for all 5-mer motifs derived using lymph node samples from three individuals (3931LN, 4014LN, and 4106LN) are shown along the diagonal. Mutability values are ranked (from lowest to highest) and color coded by their category (WRC/GYW are red, SYC/GRS are blue, WA/TW are green, and the rest are gray). Symbols indicate the mutated nucleotide (in the center of the 5-mer). Correlations between the mutabilities for all 5-mer motifs across individuals are shown in the upper (log-log scale) and lower (linear scale) triangles.

base-pair paired-end reads (8 samples) and Roche/454 GS FLX (3 samples). Details are provided in **Table 1**. These samples were originally collected and sequenced as part of three ongoing studies (manuscripts in preparation).

3.1.1. Illumina MiSeq data

Human lymph node specimens were collected under an exempt protocol approved by the Human Research Protection Program at Yale School of Medicine. Tissues were processed and RNA isolated as previously described (25). Blood samples were collected under the approval of the Personal Genome Project (26). Total RNA was immediately extracted from each blood sample and stored at −80°C until use. To carry out sequencing, mRNA was reverse transcribed into cDNA using gene-specific primers mapping to the constant region of the Ig heavy chain. Resultant cDNA was tagged with 17 nucleotide single-molecule barcodes and amplified

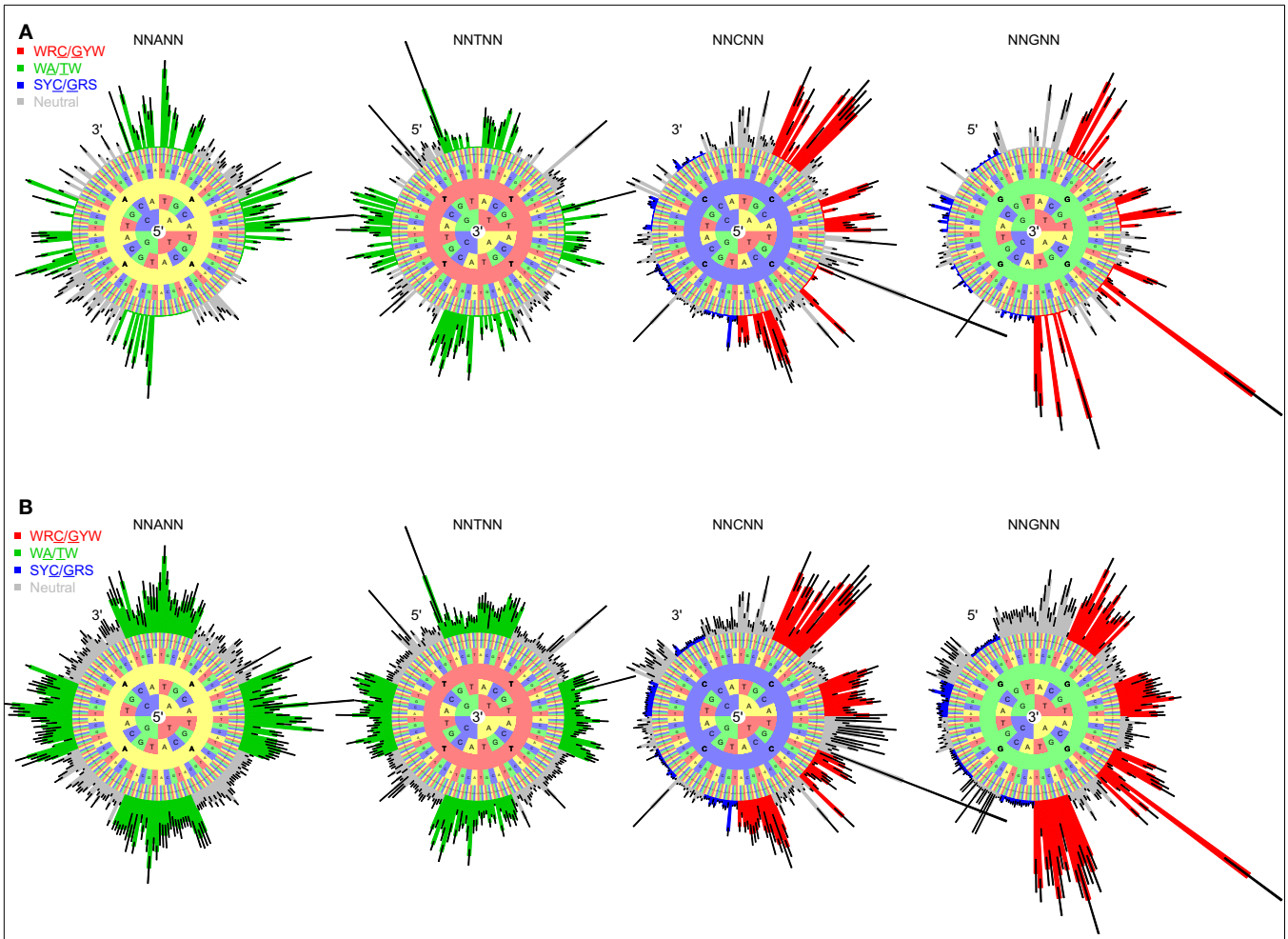


FIGURE 3 | "Hedgehog" plots for the S5F targeting model for somatic hypermutation targeting A, T, C, and G nucleotides (individual circles). **(A)** 5-mer mutabilities estimated directly from the Ig sequencing data. **(B)** The complete S5F targeting model after inferring values for missing 5-mer motifs. Bars radiating from each circle depict the mutability as a function of surrounding bases. The inner-most ring in the circle corresponds to two positions upstream (5′) of the mutated base for A and C, and two positions downstream (3′) of the mutated base for T and G. Bar colors indicate known hot/cold-spot motifs (WRC/GYW are red, WA/TW are green, SYC/GRS are blue, and "neutral" are gray). Each plot corresponds to a different mutated nucleotide. Error bars indicate confidence intervals based on the middle 50% quantiles of the mutability value distribution across the set of samples.

by PCR in a multiplex reaction using primer sets for all possible V-regions ($n = 45$) and isotype/J-regions ($n = 6$) to generate heavy chain transcripts. The amplified library was tagged with barcodes for sample multiplexing, PCR enriched, and annealed to the required Illumina clustering adapters. High-throughput 250 base-pair paired-end sequencing was performed using the Illumina MiSeq platform. Raw reads were exported without the sample barcodes and Illumina clustering adapters.

3.1.2. *Roche/454 GS FLX data*

Blood samples were collected under the approval of the Personal Genome Project (26). Total RNA was immediately extracted from each blood sample and stored at −80°C until use. Ig heavy chain mRNA were reverse-transcribed using a pool of 6 primers specific to the Ig constant regions and cDNA was amplified using 16 cycles of PCR with a pool of 46 V-region-specific primers and 6 nested constant region primers. Following ligation of 454-compatible sequencing adapters, the expected heavy chain V gene fragments were purified using PAGE. Each sample was uniquely barcoded during the ligation process, allowing subsequent mixing of all the samples into one common reaction sample (performed independently for each replicate run). Emulsion PCR and 454 GS FLX sequencing were performed directly at the 454 Life Sciences facility according to the manufacturer's standard protocols.

3.2. SEQUENCING DATA PRE-PROCESSING

Raw sequencing reads were filtered in several steps to identify and remove low-quality sequences. Conservative thresholds were applied in all cases to increase the reliability of the resulting mutation calls, at the potential expense of excluding some real mutations. Pre-processing was carried out using the Repertoire

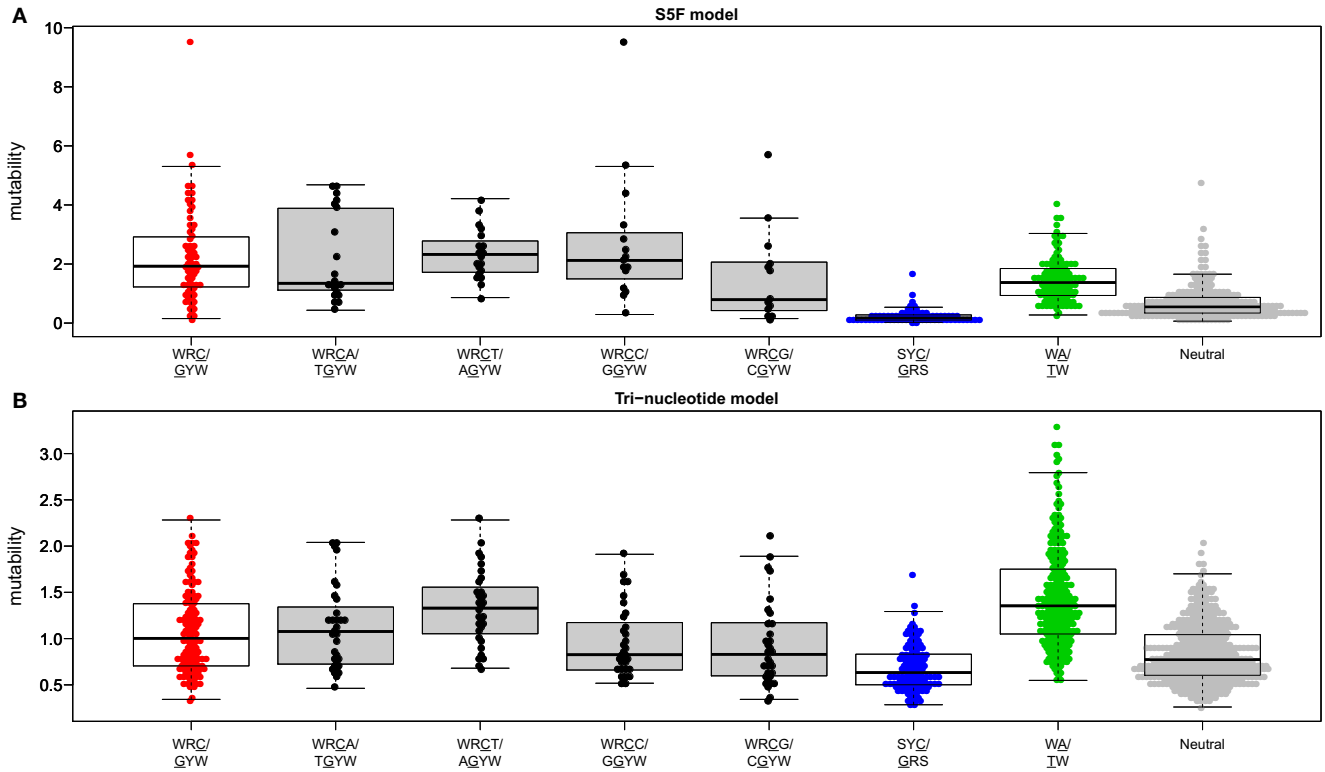


FIGURE 4 | The hierarchy of hot/cold-spot motifs. Mutability values predicted by the **(A)** S5F model or **(B)** the tri-nucleotide model of Shapiro et al. were grouped by the type of motif. Boxes borders correspond to the first and fourth quartiles while the horizontal bar inside the box corresponds to the median of the distribution. Hot/cold spot motif groups are plotted using the same colors as in **Figure 3B** with open boxes. WRCN/NGYW motifs are plotted using filled boxes.

Sequencing Toolkit (pRESTO) (http://clip.med.yale.edu/pRESTO, manuscript in preparation), and involved:

- Quality control
  1. Removal of low-quality reads (mean Phred quality score <20).
  2. Removal of reads where the primer could not be identified or had a poor alignment score (mismatch rate greater than 0.1).
  3. For the MiSeq data, sets of sequences with identical molecular IDs (corresponding to the same mRNA molecule) were identified. Sets were collapsed into one consensus sequence per set, after discarding those having a mean mismatch rate across all positions >0.2.
  4. For the MiSeq data, the two paired-end reads were assembled into a complete Ig sequence.
  5. Removal of sequences that do not appear in a single sample at least twice.
- Assignment of germline V(D)J segments for each of the Ig sequences: initial V(D)J assignments for each sequence were obtained using IMGT/HighV-QUEST (27). Using these assignments, non-mutated sequences were identified and a V segment germline repertoire for each individual was determined as the set of: (1) V genes that composed at least 0.1% of the sequences, and (2) V gene alleles that composed at least 10% of the assignments to that V gene. Ig sequences that were initially assigned V segments not included in this germline repertoire were then re-assigned to the closest present V segment based on the Hamming distance.
- Removal of non-functional sequences due to the occurrence of a stop codon or/and a reading frame shift between the V gene and the J gene.
- Removal of sequences with more than 30 mutations and masking (replacement with Ns) of positions with Phred quality scores <20.
- Removal of mutations in codons that had more than one mutation, as it is usually not possible to infer the order in which the mutations occurred (and thus the micro-sequence context of the mutations is unknown).
- Identification of clonally related sequences: a two-step approach was applied to identify sequences that were part of a B cell clone (i.e., related through descent from a common ancestor). First, the sequences were divided into groups based on equivalence of their V-gene assignment, J-gene assignment, and the number of nucleotides in their junction. Second, clones were defined within each of these groups as the collection of sequences with junction regions that differed from one sequence to any of the others by no more than three point mutations.

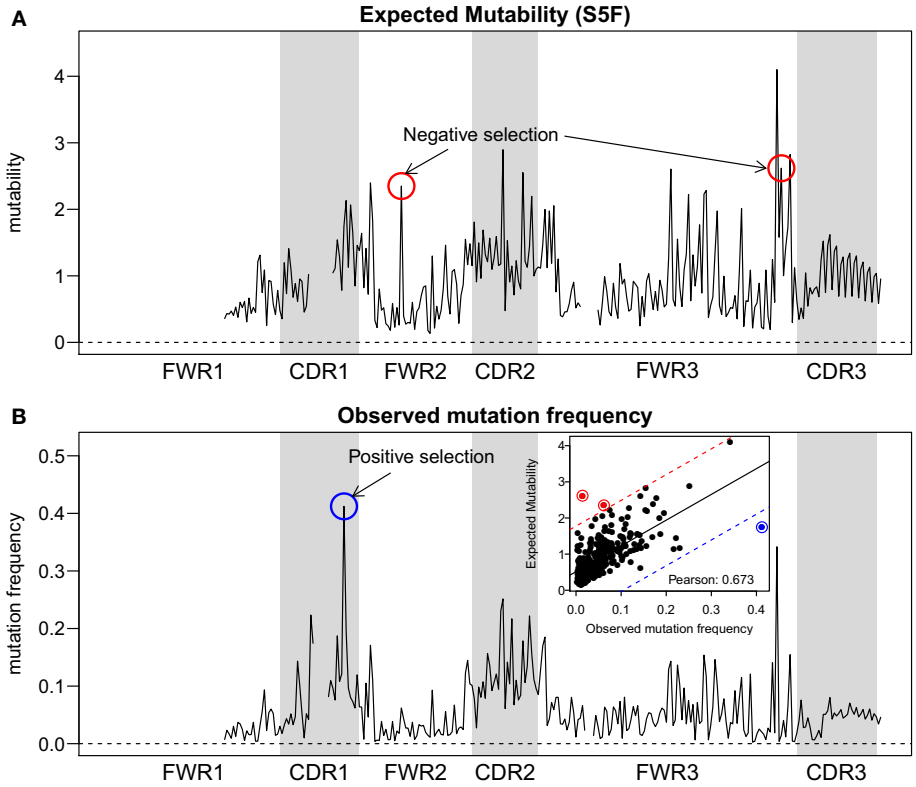


FIGURE 5 | Comparison between expected and observed somatic hypermutation targeting. **(A)** The predicted mutability from the S5F model and **(B)** the observed mutation frequency from sample 3931LN (averaged over all clones) for each position in the Ig sequence (IMGT-aligned along the *x*-axis). The correlation across positions (points) is shown in the inset of **(B)**. Two positions with evidence of negative selection (red circles) and one position with positive selection (blue circles) are indicated. The threshold for calling a position with significant selection was set to 3 SD away from the linear regression line (shown as a solid line in the inset, with thresholds plotted as dashed lines).

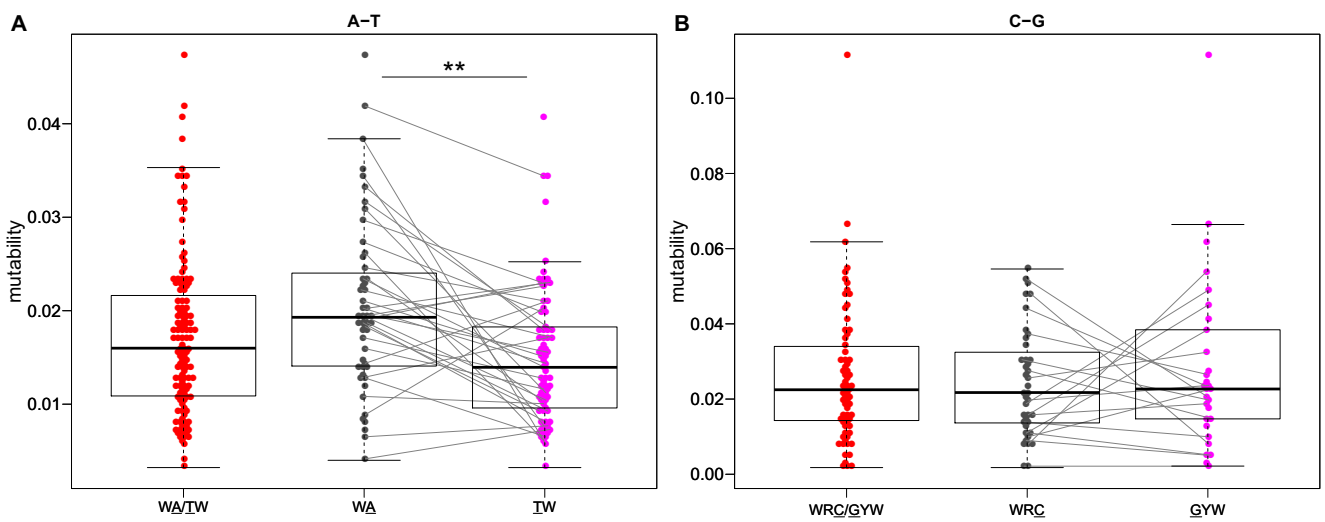


FIGURE 6 | Somatic hypermutation targeting at C/G, but not A/T, hot-spots is strand symmetric. Mutability values directly estimated by the S5F model (**Figure 3**) for **(A)** WA and **(B)** WRC "hot-spot" motifs are compared between different strands. Lines connect reverse-complement motifs for cases where both could be directly estimated from the data. $^{**}P < 5 \times 10^{-4}$ by a paired Mann–Whitney–Wilcoxon test.

The threshold of three was determined after manual inspection of the mutation patterns in resulting clones identified through building lineage trees.

4. DISCUSSION

We have constructed new SHM targeting and substitution models using a collection of more than 800,000 synonymous mutations from next-generation Ig sequencing studies. The exclusive use of synonymous mutations allowed us to include mutations from functional Ig sequences without the biasing influence of selection. The large size of the resulting mutation data set allowed us to model targeting and substitution dependencies on the mutating base as well as on two bases upstream and downstream of the mutation. The resulting "S5F" models validate, and also help refine, previously defined SHM hot and cold spots. **Figure 4** shows how the classic WRCY/RGYW hot-spot excludes some highly mutable WRCA/TGYW motifs, implying that, as proposed by Rogozin and Diaz (15), WRCH/DGYW could be a better predictor of mutation. However, while the most mutable WRCA/TGYW motifs are even hotter than WRCY/RGYW, others are comparable to neutral motifs. This high variance demonstrates the importance of including higher order dependencies, as we have done.

It has been suggested that nucleotide substitution profiles are also dependent on the micro-sequence context of the mutating base (8, 17). We confirm that the substitution profiles at all nucleotides are highly dependent on neighboring bases and these dependencies are conserved across individuals. Interestingly, the fact that substitution rates depend on surrounding bases may resemble the situation in meiotic mutations as was suggested in the past (9). The ability of the S5F models to estimate mutability and substitution at each of the 1024 DNA 5-mer motifs will allow for detailed, quantitative comparison of SHM with other mutation processes.

A potential source of error in the approach taken here is the existence of novel polymorphisms among the seven individuals studied (**Table 1**). Since mutation detection depends on comparison with known V and J segments that are part of the IMGT repertoire, undetected polymorphisms will look like mutations. However, any effect on the S5F model is expected to be small relative to the estimated confidence intervals. Based on a new statistical tool to detect novel germline alleles from high-throughput sequencing data (manuscript in preparation), the magnitude of this effect was estimated to be less than ∼1% of the sequences and less than ∼0.1% of the mutations used for the current analysis. The S5F mutability and substitution models presented here were developed using human heavy chain data, and thus may not be valid for light chains or mouse sequences. Given the large amount of sequencing data becoming available, it may be possible to extend the proposed approach to model 7-mers instead of 5-mers. However, even with 5-mers, the values for some motifs had to be inferred because of the limited diversity in germline repertoires. It will be important to estimate the quality of these inferences experimentally. Future experiments might be designed to enrich for non-productively rearranged Ig sequences which could then be sequenced using high-throughput technologies. Since mutations in these sequences are (presumably) not subject to selection, they provide a way to independently estimate substitution profiles and mutabilities for at least some of the motifs inferred in the S5F model. It will be important to confirm that the mutation process operating on these non-productive sequences is equivalent to the process at the productive alleles. This uncertainty is one reason why only productively rearranged Ig sequences were included in the current model.

The targeting and substitution models developed here provide a quantitative description of SHM in the absence of selection, and thus provide an important background for statistical analysis of SHM patterns in experimental data. For example, such models play an important role in quantifying antigen-driven selection in Ig sequences (11, 20), and we have now made the S5F model available as an option on our website for quantifying selection (http://clip.med.yale.edu/baseline). When combined with high-throughput sequencing, it should now be possible to quantify selection for each position of the Ig sequence independently and link these values back to the physical structure of the protein. Following the approach of Brard and Guguen (28), these models could also be incorporated into methods for building lineage trees of B cell clones (29), thus helping to provide insight into the underlying population dynamics of adaptive immunity. The model and method source code are made available at http://clip.med.yale.edu/SHM.

ACKNOWLEDGMENTS

We thank the Yale High Performance Computing Center (funded by NIH grant: RR19895) for use of their computing resources. Funding: this work was partially supported by NIH R03AI092379. The work of Jason A. Vander Heiden, Daniel Gadala-Maria and Namita Gupta were supported in part by NIH Grant T15 LM07056 from the National Library of Medicine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/journal/10.3389/fimmu.2013.00358/abstract

REFERENCES

1. Chahwan R, Edelmann W, Scharff MD, Roa S. AIDing antibody diversity by error-prone mismatch repair. *Semin Immunol* (2012) **24**(4):293–300. doi:10.1016/j.smim.2012.05.005
2. Peled JU, Kuang FL, Iglesias-Ussel MD, Roa S, Kalis SL, Goodman MF, et al. The biochemistry of somatic hypermutation. *Annu Rev Immunol* (2008) **26**(1):481–511. doi:10.1146/annurev.immunol.26.021607.090236
3. McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc Natl Acad Sci U S A* (1984) **81**(10):3180–4. doi:10.1073/pnas.81.10.3180
4. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol* (2003) **171**(9):4639–49.
5. Betz AG, Rada C, Pannell R, Milstein C, Neuberger MS. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc Natl Acad Sci U S A* (1993) **90**(6):2385–8. doi:10.1073/pnas.90.6.2385
6. Shapiro GS, Aviszus K, Ikle D, Wysocki LJ. Predicting regional mutability in antibody v genes based solely on di- and trinucleotide sequence composition. *J Immunol* (1999) **163**(1):259–68.
7. Smith DS, Creadon G, Jena PK, Portanova JP, Kotzin BL, Wysocki LJ. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive b cells. *J Immunol* (1996) **156**:2642–52.
8. Cowell LG, Kepler TB. The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric

and distinct from that under germline mutation. *J Immunol* (2000) **164**(4):1971–6.
9. Oprea M, Cowell LG, Kepler TB. The targeting of somatic hypermutation closely resembles that of meiotic mutation. *J Immunol* (2001) **166**(2):892–9.
10. Shapiro GS, Ellison MC, Wysocki LJ. Sequence-specific targeting of two bases on both DNA strands by the somatic hypermutation mechanism. *Mol Immunol* (2003) **40**(5):287–95. doi:10.1016/S0161-5890(03)00101-9
11. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) **39**(Suppl 2):W499–504. doi:10.1093/nar/gkr413
12. Shlomchik MJ, Watts P, Weigert MG, Litwin S. Clone: a Monte-Carlo computer simulation of b cell clonal expansion, somatic mutation, and antigen-driven selection. *Curr Top Microbiol Immunol* (1998) **229**:173–97. doi:10.1007/978-3-642-71984-4_13
13. MacCarthy T, Kalis SL, Roa S, Pham P, Goodman MF, Scharff MD, et al. V-region mutation in vitro, in vivo, and in silico reveal the importance of the enzymatic properties of AID and the sequence environment. *Proc Natl Acad Sci U S A* (2009) **106**(21):8629–34. doi:10.1073/pnas.0903803106
14. Spencer J, Dunn M, Dunn-Walters DK. Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. *J Immunol* (1999) **162**(11):6596–601.
15. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol* (2004) **172**(6):3382–4.
16. Bransteitter R, Pham P, Calabrese P, Goodman MF. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J Biol Chem* (2004) **279**(49):51612–21. doi:10.1074/jbc.M408135200
17. Spencer J, Dunn-Walters DK. Hypermutation at A-T base pairs: the A nucleotide replacement spectrum is affected by adjacent nucleotides and there is no reverse complementarity of sequences flanking mutated A and T nucleotides. *J Immunol* (2005) **175**(8):5170–7.
18. Bose B, Sinha S. Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection. *Immunology* (2005) **116**(2):172–83. doi:10.1111/j.1365-2567.2005.02208.x
19. Cohen RM, Kleinstein SH, Louzoun Y. Somatic hypermutation targeting is influenced by location within the immunoglobulin Z region. *Mol Immunol* (2011) **48**(12-13):1477–83. doi:10.1016/j.molimm.2011.04.002
20. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) **40**(17):e134. doi:10.1093/nar/gks457
21. Correa J. *Interval Estimation of the Parameters of the Multinomial Distribution*. Statistics on the Internet (2001). Available from: interstat.statjournals.net.
22. Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc B* (1951) **13**(2):238–41.
23. Kunik V, Ashkenazi S, Ofran Y. Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res* (2012) **40**:W521–4. doi:10.1093/nar/gks480
24. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic mutation hotspots correlate with DNA polymerase error spectrum. *Nat Immunol* (2001) **2**(6):530–6. doi:10.1038/88732
25. Willis SN, Mallozzi SS, Rodig SJ, Cronk KM, McArdel SL, Caron T, et al. The microenvironment of germ cell tumors harbors a prominent antigen-driven humoral response. *J Immunol* (2009) **182**(5):3310–7. doi:10.4049/jimmunol.0803424
26. Nirantar SR, Ghadessy FJ. Compartmentalized linkage of genes encoding interacting protein pairs. *Proteomics* (2011) **11**(7):1335–9. doi:10.1002/pmic.201000643
27. Lefranc M-P, Pommi C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Dev Comp Immunol* (2003) **27**(1):55–77. doi:10.1016/S0145-305X(02)00039-3
28. Brard J, Guguen L. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Syst Biol* (2012) **61**(3):510–21. doi:10.1093/sysbio/sys024
29. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) **338**(12):67–74. doi:10.1016/j.jim.2008.06.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2013; accepted: 22 October 2013; published online: 15 November 2013.
*Citation: Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JNH, O'Connor KC, Hafler DA, Laserson U, Vigneault F and Kleinstein SH (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front. Immunol. **4**:358. doi: 10.3389/fimmu.2013.00358*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Yaari, Vander Heiden, Uduman, Gadala-Maria, Gupta, Stern, O'Connor, Hafler, Laserson, Vigneault and Kleinstein. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 November 2013
doi: 10.3389/fimmu.2013.00357

Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures

Gregory W. Schwartz [1] ***and Uri Hershberg*** [1,2] *******

[1] *Systems Immunology Laboratory, School of Biomedical Engineering, Science, and Health Systems, Drexel University, Philadelphia, PA, USA*
[2] *Department of Microbiology and Immunology, College of Medicine, Drexel University, Philadelphia, PA, USA*

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Andrew M. Collins, University of New South Wales, Australia
Nicole Wittenbrink, Humboldt University Berlin, Germany

****Correspondence:***
Uri Hershberg, Systems Immunology Laboratory, School of Biomedical Engineering, Science, and Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
e-mail: uri.hershberg@drexel.edu

The diversity of the immune repertoire is important for the adaptive immune system's ability to detect pathogens. Much of this diversity is generated in two steps, first through the recombination of germline gene segments and second through hypermutation during an immune response. While both steps are to some extent based on the germline level repertoire of genes, the final structure and selection of specific receptors is at the somatic level. How germline diversity and selection relate to somatic diversity and selection has not been clear. To investigate how germline diversity relates to somatic diversity and selection, we considered the published repertoire of Ig heavy chain V genes taken from the blood of 12 individuals, post-vaccination against influenza, sequenced by 454 high-throughput sequencing. We here show that when we consider individual amino acid positions in the heavy chain V gene sequence, there exists a strong correlation between the diversity of the germline repertoire at a position and the number of B cell clones that change amino acids at that position. At the same time, we find that the diversity of amino acids used in the mutated positions is greater than in the germline, albeit still correlated to germline diversity. From these findings, we propose that while germline diversity and germline amino acid usage at a given position do not fully specify the amino acid mutant needed to promote survival of specific clones, germline diversity at a given position is a good indicator for the potential to survive after somatic mutation at that position. We would therefore suggest that germline diversity at each specific position is the better *a priori* model for the effects of somatic mutation and selection, than simply the division into complementarity determining and framework regions.

Keywords: B cells, somatic hypermutation, selection, diversity, evolution

1. INTRODUCTION

The adaptive immune system's ability to react to disease is based on the diversity of its immune repertoire. In the case of B cells, this diversity is generated in two rounds: the recombination of germline gene segments (V, D, and J for heavy chains, V and J for light chains) to create the B cell receptor (BCR), (1–3) as well as somatic hypermutation during an immune response (4–6). In both cases, these diversification processes are coupled with stringent somatic selection based on the binding affinity of the BCR (7). Thus, while the initial state of the BCR is at least somewhat based on an individual's germline genes, the final structure of specific BCR mutants is based on somatic selection processes related to the binding affinity of the BCR. It remains unclear how selection and diversity at the germline level relate to selection at the somatic level. In this analysis, we demonstrate a link between the diversity and selection at the germline and somatic levels for V genes.

The germline genes encoding the different regions of the BCR are themselves diverse, even before considering the diversification produced from the recombination of different gene segments. Specifically in V genes, this diversity is non-uniformly spread across the gene sequence. Some positions always utilize the same amino acid in all V genes while others can utilize many different amino acids. This differential diversity is considered an indicator of the functional role of each position in the eventual tertiary structure of the receptor. Variable regions of the V gene sequences, called complementarity determining regions (CDR), are thought to be those that encode regions which interact with antigens, while the more invariant positions, called framework regions (FR), are proposed to be involved in the backbone of the receptor (8). It has been generally thought that somatic selection segregates along these two regions. Positive selection is thought to occur in the CDR, while mutations in the FR were mostly debilitating to affinity and lethal to the cell (5). It is now clear that this segregation is not strictly true – positively selected key mutations can be found in the FR (9) and negative selection can be seen in mutations throughout the sequence (10).

Previously, the diversity measurements of the receptors were based on differing diversity indexes with varying appropriateness and on partial sets of germline and mutant sequences (8, 11). We directly measured the "true" diversity of light chain V genes and heavy chain V genes (V_H) based on the entire known BCR germline repertoire as found in the IMGT Ig gene database (12, 13). We demonstrated that the pattern of diversity in all V genes

is non-uniform, with most positions showing a low level of diversity (2–5 relevant amino acids) and a few exhibiting higher levels of diversity (up to 10 relevant amino acids). If we rank all the positions in the sequence by their diversity, we can explicitly show that while the CDR is enriched for high diversity positions, many CDR positions have diversities as low as those found in FR and some FR positions have quite a high diversity (12). We previously suggested that it is the diversity of positions, not solely their association with the contiguous CDR or FR positions, which determines their functional role and the consequence of mutation.

The diversity of positions in the germline repertoire of V genes is the result of evolutionary selection of individuals and their progeny. The process of affinity maturation is based on somatic mutation and selection. It has thus far been unclear how these two processes of selection are related and if they can be connected at the V gene sequence level. To study this possible relationship, we considered a published dataset of ∼17,000 recombined BCR V_H gene sequences from 12 individuals (14). We divided sequences by their clonotypes, identifying the clonal origin for each recombined sequence. In this way we could now count how many times each position was mutated in the repertoire. Comparing the number of individual times a position was mutated to its germline diversity (12), we found that while synonymous mutations were evenly spread across all positions, there was a clear positive correlation between the number of times a position had a mutated amino acid and that position's diversity in the germline repertoire. From this we conclude that the diversity at the germline level is an indication for the potential for somatic harm as a result of mutation. The diversity of each specific position is a more direct measure of the functional consequence of mutation and selection at the somatic level than a mere division into CDR and FR.

2. MATERIALS AND METHODS

2.1. SEQUENCES ANALYZED

We analyzed the amino acid and nucleotide sequences of *Homo sapiens* BCR recombined V_H genes (14). The sequences came from twelve healthy individuals post-vaccination against influenza (14). The individuals came from two age cohorts: 6 young (age range 19–45) and 6 old (age range 70–89). Sequences were acquired at days 0, 7, and 28 post-vaccination and included both IgG, IgM, and IgA class switched receptors. We divided the sequences into clones by fully aligning their nucleotide sequences to the germline V, D, and J genes from the IMGT Ig database (13). All sequences that shared the same germline source (V, J, and CDR3 length) were considered to be from the same clone. We filtered out sequences with ≥30 nt point mutations from the germline. This alignment resulted in the identification of 17,553 sequences divided into 9482 clones. Due to sequencing issues in the original dataset, we only analyzed the sequences from position 25 and on. IMGT numbering leaves gaps in order to remain consistent with all V genes. Also, the length of V genes is not always identical. Therefore, we only calculated germline diversity for amino acid positions 25 − 30, 35 − 59, 63 − 72, and 74 − 106, leaving us with 74 positions in the analysis. These positions were verified for adequate sampling by the use of rarefaction curves at each position (15). We considered a position viable if more than 99% of the curve consisted of a richness of ≥95% of the height of the curve (12). These curves rule out the possibility of having too many gaps in the germline repertoire.

We calculated the germline diversity per amino acid position using BCR V_H genes collected from IMGT as in Ref. (12). We filtered out non-functional, partial, and duplicate sequences for the analysis. All sequences were numbered according to the IMGT unique numbering system based off of the universal alignment provided by IMGT (13). We defined CDR and FR positions as in Ref. (16).

2.2. DIVERSITY MEASURES

We measured the diversity of amino acids per position as in Ref. (12) with an order of diversity equal to 1. The process of measuring diversity is dependent on the order, or "Hill number" (17), we use during calculations. While measuring the effective number of species, the order affects the influence of the sample abundances. An order of 0 does not consider abundances, thus all types are considered equally (this is equivalent to the number of different types, also called "richness"). An order gives greater weight to rare species, while an order >1 gives greater weight to common species. When the order is 1, the effective diversity is determined without any bias (18). We previously described the result of analyzing the diversity of the different amino acid positions in the V gene germline repertoire at different orders of diversity (12). We decided here to focus on the order of 1 as we found no *a priori* reason to bias toward either the more commonly used amino acids at each position or toward the rare amino acids.

At each position p, the number of amino acids at that position was N_p and the richness of the amino acids at that position was R_p.

The measure of diversity used for these positions was "true" diversity ${}^qD_p$, where

$$ {}^qD_p \equiv \left(\sum_{i=1}^{R_p} p_{i_p}^{q} \right)^{(1/1-q)} \tag{1} $$

and q is the order of diversity and p_i is the frequency of amino acid i (17, 18). At $q = 1$, equation (1) does not exist, however the limit as q approaches 1 is

$$ {}^1D_p \equiv exp\left(-\sum_{i=1}^{R_p} p_{i_p} \ln p_{i_p} \right) \tag{2} $$

2.3. DEFINITION OF POSITION CATEGORIES

For every amino acid position, we counted – across all clones from any time point and person – the number of times a position changed amino acids and how many times that position maintained its amino acid from the germline. If a position was found to change into several amino acids in a single clone, that position was counted once for each different amino acid. The cases where amino acids were maintained relative to the germline were in some cases further divided into non-mutated and synonymous mutations. The amino acids collected in each category (changed, maintained, or synonymous mutation) were then further analyzed for their diversity and amino acid composition tendencies.

2.4. CORRELATIONS OF DIVERSITY IN GERMLINE POSITIONS VERSUS CHANGED OR MAINTAINED AMINO ACID POSITION CATEGORIES

Using a two sided Spearman's rank correlation test, we assessed the correlation of germline diversity of human V_H genes, as calculated in Ref. (12), with the counts and diversities of the three categories (changed amino acid, maintained amino acid, and synonymous mutation) described above.

2.5. AMINO ACID USAGE ANALYSIS

We assessed if position categories were biased toward specific amino acid usage types. Following our definitions of Ig relevant amino acid categorization by hydrophobicity and tendency to be found on the surface of the receptor (16, 19), we categorized amino acids as hydrophobic (IVLFCMW), neutral (AGTSYPH), and hydrophilic (NDQEKR) (12). We then categorized a position by how biased that position was to using amino acids from only one of these categories. If a position used only amino acids from one category, that position was considered to be of that type (i.e., a hydrophobic, a neutral, or a hydrophilic position). If the position had both neutral and one other category of amino acids, that position would be considered a "weak" version of that category (i.e., weak hydrophobic or weak hydrophilic). If there were amino acids in all categories, then that position was considered indeterminate. In all instances, if a position had a single amino acid in one category, and three or more in another category, the single amino acid category was ignored (12).

3. RESULTS

3.1. CORRELATION OF GERMLINE DIVERSITY TO NUMBER OF CHANGED AMINO ACIDS PER POSITION

When comparing the germline diversity at each position – as calculated from the prototypical IMGT database (12) – and the number of unique changed amino acids at each position, we find that these two properties are highly positively correlated ($\rho = 0.710$, $p = 1.41 \times 10^{-12}$). This correlation holds true if we consider CDR ($\rho = 0.676$, $p = 2.18 \times 10^{-4}$) and FR ($\rho = 0.668$, $p = 2.17 \times 10^{-7}$) positions separately and if we consider all positions as a whole (**Figure 1**). While this correlation is monotonic, it is by no means strictly linear as the linear model explains only ∼43% of variation in the extent of amino acid exchanges at the different positions

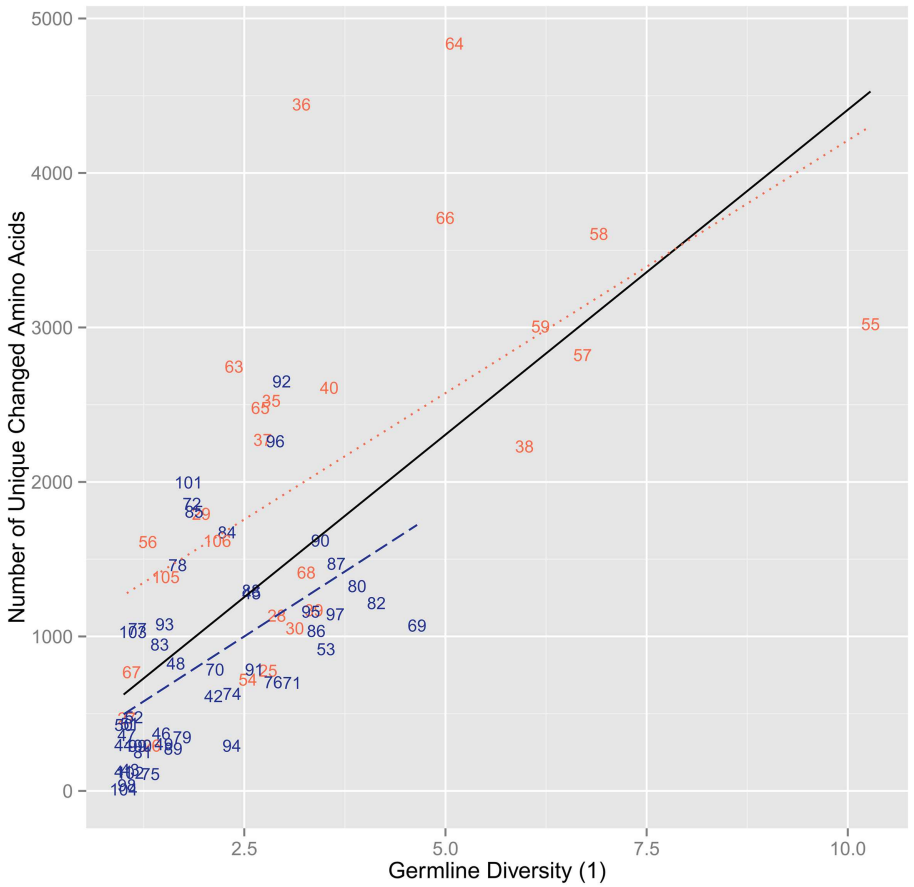


FIGURE 1 | The number of unique changed amino acids versus the diversity of order 1 of the germline sequences per position. The points are labeled by their IMGT sequence position number and if they are found in the CDR (red) or the FR (blue). The lines represent linear regressions for all positions (black, $r^2 = 0.433$), for FR positions (dashed blue, $r^2 = 0.289$), and for CDR positions (dotted red, $r^2 = 0.349$). We found a significant positive correlation for all positions ($\rho = 0.710$, $p = 1.41 \times 10^{-12}$). By correlating the positions based on FR and CDR, we found a significant positive correlation for both FR ($\rho = 0.668$, $p = 2.17 \times 10^{-7}$) and CDR ($\rho = 0.676$, $p = 2.18 \times 10^{-4}$).

($r^2 = 0.433$). Interestingly, while in general the CDR positions with similar diversity have more changed amino acids than most FR positions of similar germline diversity and the linear fits to CDR and FR are distinct, FR and CDR positions are not clearly separated in this plane (**Figure 1**).

The analysis with synonymous mutations shows no correlation ($r^2 = 1.87 \times 10^{-3}$, $\rho = 0.232$, $p = 0.0468$) and similar mutation levels across the entire range of germline position diversities and no difference between CDR ($r^2 = 0.0322, \rho = -0.223, p = 0.273$) and FR ($r^2 = 0.0127$, $\rho = 0.289$, $p = 0.0466$) positions (**Figure 2**). The results found using the diversity of the germline repertoire were at the whole repertoire level. No division into certain germlines was necessary and so the possibility for misidentification of the germlines by IMGT would have little to no impact on the diversity at the repertoire level. Moreover, when splitting up the analysis of clones by the germline they aligned with, there was no real difference in between different germlines and at the repertoire level (results now shown).

3.2. CORRELATION OF GERMLINE DIVERSITY TO CHANGED OR MAINTAINED DIVERSITY PER POSITION

We next looked to see how the actual amino acid diversity of the mutant repertoire at the different positions related to the germline diversity. We found that the maintained positions had a diversity that was essentially identical to that found in the IMGT based germline repertoire ($r^2 = 0.947, \rho = 0.961, p = 7.52 \times 10^4 2$) (**Figure 3A**). In the changed positions a more complex pattern emerges. While overall we find again that there is a positive correlation between germline diversity and the diversity of the changed amino acids ($r^2 = 0.284$, $\rho = 0.359$, $p = 1.70 \times 10^3$), the range of diversity is much greater in the changed positions (**Figure 3B**). This greater range of diversity is present in both CDR ($r^2 = 0.419$, $\rho = 0.580$, $p = 2.27 \times 10^3$) and FR ($r^2 = 0.0347$, $\rho = 0.132$, $p = 0.371$). However, when the FR is considered on its own this leads to a lack of significant correlation with germline diversity.

3.3. CHANGES IN AMINO ACID USAGE PATTERN

We categorized the amino acid usage patterns for each position. We found in the maintained amino acid positions the biases toward using specific amino acid types are maintained. This was especially true for the positions in the germline that had stricter categories of amino acids usage bias. 13 out of 14 hydrophobic positions, 17 out of 19 neutral positions, and 8 out of 10 hydrophilic positions retain the same bias in the maintained positions as in the germline (**Table 1**). The positions with the more intermediate

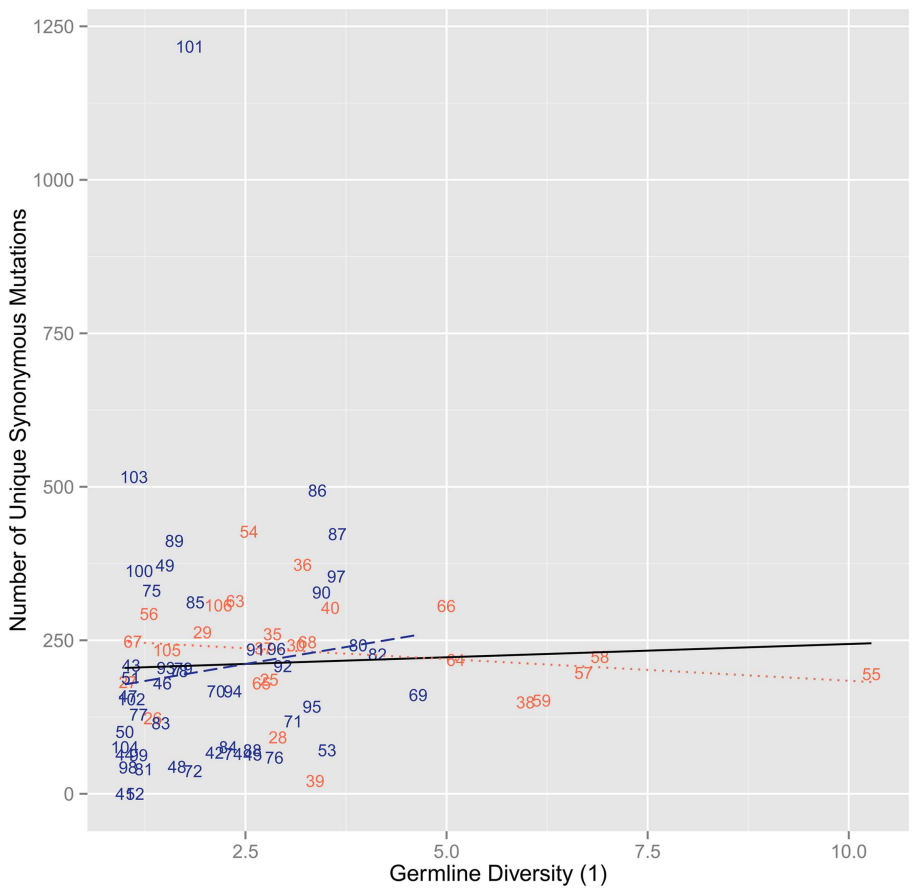


FIGURE 2 | The number of synonymous mutations versus the diversity of order 1 of the germline sequences per position. We found no positive correlation with a flat trend ($r^2 = 1.87 \times 10^{-3}$, $\rho = 0.232$, $p = 0.0468$). When splitting by region, we found no correlation for FR ($r^2 = 0.0127$, $\rho = 0.289$, $p = 0.0466$) and CDR ($r^2 = 0.0322$, $\rho = -0.223$, $p = 0.273$).

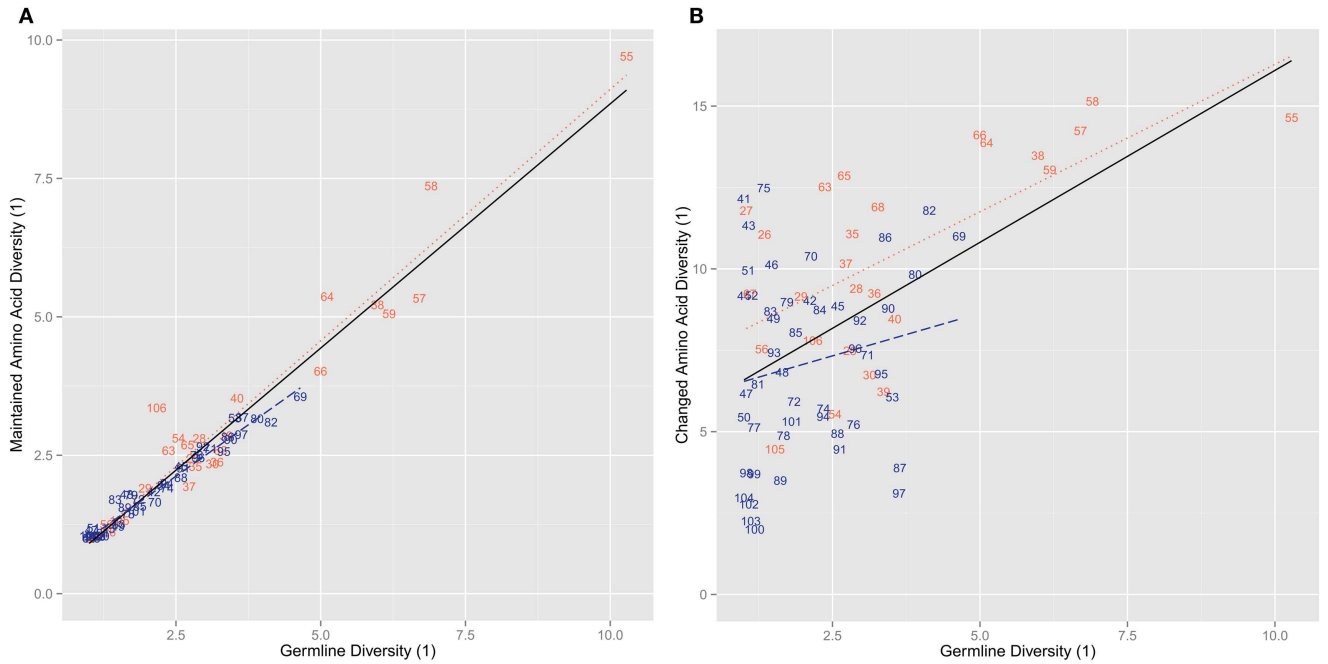


FIGURE 3 | The diversity of order 1 of the amino acids versus the diversity of order 1 for the germline sequences per position. **(A)** The diversity of order 1 amino acid usage in amino acids positions. We found a significant positive correlation ($r^2 = 0.947$, $\rho = 0.961$, $p = 7.52 \times 10^{-42}$). We also found positive correlations when the positions were split by region for FR ($r^2 = 0.970$, $\rho = 0.968$, $p = 3.67 \times 10^{-29}$) and CDR ($r^2 = 0.937$, $\rho = 0.880$, $p = 1.76 \times 10^{-6}$). **(B)** The diversity of order 1 amino acid usage of unique changed amino acids. We found a positive correlation ($r^2 = 0.284$, $\rho = 0.359$, $p = 1.70 \times 10^{-3}$). When splitting the positions by regions, however, FR did not correlate with the germline diversities ($r^2 = 0.0347$, $\rho = 0.132$, $p = 0.371$) while CDR did have a positive correlation ($r^2 = 0.419$, $\rho = 0.580$, $p = 2.27 \times 10^{-3}$).

Table 1 | Number of positions for each amino acid usage bias.

Category	Germline	Maintained	Changed
Hydrophobic	14	16: (13, 3, 0, 0, 0, 0)	7: (6, 1, 0, 0, 0, 0)
Weak hydrophobic	12	0: (0, 0, 0, 0, 0, 0)	8: (5, 2, 1, 0, 0, 0)
Neutral	19	27: (0, 6, 17, 4, 0, 0)	2: (0, 0, 2, 0, 0, 0)
Weak hydrophilic	13	3: (0, 0, 1, 2, 0, 0)	5: (0, 0, 1, 3, 1, 0)
Hydrophilic	11	10: (0, 0, 0, 1, 8, 1)	0: (0, 0, 0, 0, 0, 0)
Indeterminate	5	18: (1, 3, 1, 6, 3, 4)	52: (3, 9, 15, 10, 10, 5)

The numbers in parentheses after the colon signify the classifications of the respective positions in the germline repertoire: hydrophobic, weak hydrophobic, neutral, weak hydrophilic, hydrophilic, indeterminate.

biases (weak hydrophobic and weak hydrophilic) in the germline did not adhere as strictly to the same bias category but tend to have changed to one of the neighboring biases. Weak hydrophobic becomes either hydrophobic or neutral. Weak hydrophilic becomes either hydrophilic or neutral (**Table 1**). Looking now at the changed position we see that biases change much more (**Figure 4**). Most positions simply become indeterminate (i.e., have no clear bias). However, it is interesting to note that those positions that do have some bias exhibit either exactly the same bias as they have in the germline repertoire or one that is similar (**Table 1**).

4. DISCUSSION

The specificity of B cell and T cell receptors, while based on genes in the germline, is ultimately not of the germline template. Due to the imprecise nature of V(D)J recombination and, in B cells, somatic mutation, the final affinity of each immune receptor is neither inherited nor heritable. For this reason it is difficult to assess how germline diversity and its selection relate to selection during an immune response and specifically how they relate to the anticipated outcome of somatic mutation during an immune response. We have previously shown that the diversity of the germline V gene repertoire can be characterized by looking at the amino acid diversity of individual positions in the V gene sequence (12). The distribution of diversity is non-uniform with most, but not all highly diverse positions being found in the CDR. Furthermore, different positions show different biases toward the use of hydrophobic or hydrophilic amino acids (12). To contrast this picture of germline diversity with somatic changes, we have taken a published sample of the human peripheral B cell repertoire following influenza vaccination. We divided all of the sequences in this dataset into their respective clones and counted the number of times each position in the V gene sequence changed or maintained the amino acid found in the germline origin of its clone. By doing so, we could compare for each position how it contributes to repertoire diversity and its selection when changed from its germline and when it remained the same. Analyzing the maintained positions and their diversity allows us to ask to what extent clonal shift

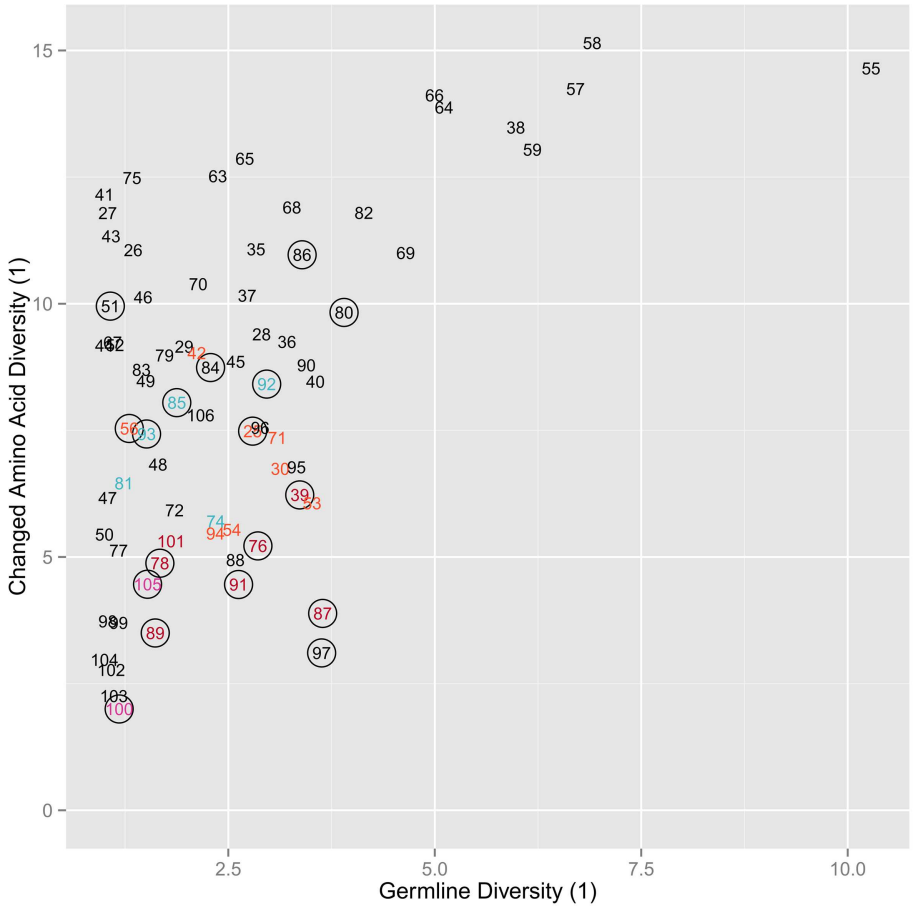


FIGURE 4 | Copy of Figure 3B annotated for amino acid usage bias. The coloring for the labels signifies whether that position is hydrophobic (red), weak hydrophobic (light red), neutral (purple), weak hydrophilic (light blue), hydrophilic (blue), or indeterminate (black), while a circle around the position indicates that the position shares the same hydrophobicity bias as the equivalent germline position. All amino acids are changed.

changes the diversity of the repertoire from the germline while analyzing the changed positions describes the effects of selection.

Starting with the maintained positions, we see that the germline diversity exhibited in the prototypic repertoire in the IMGT database, which does not assume any specific biases in V_H usage, is recapitulated in even the small and clonally shifted snapshot of the immune repertoire analyzed here. We find a significant linear correlation between germline diversity and that of positions with maintained amino acids (**Figure 3A**) and also clear conservation of amino acid usage biases (**Table 1**). Thus, despite the fact that we analyzed ∼1000 clones per person (out of potentially 10^{11} clones) with a significant shift toward certain V_H genes, we still identify more or less exactly the same diversity and amino acid usage as described by the IMGT database. This suggests to us that clonal shift does not change the make up of amino acid diversity in the B cell repertoire. Furthermore, the existing IMGT database of human V genes represents this positional diversity well.

With regard to the changed positions, we find that there is a significant positive correlation between the level of diversity in the germline at a specific position and the survival of clones with changed amino acids at that position (**Figure 1**). Such a correlation suggests that there is a relationship between the tendency to diversify a position at the germline level over evolutionary time and the likelihood of mutants at those positions to survive somatic mutation and selection. We do not find any kind of correlation between germline diversity and synonymous mutation level (**Figure 2**). For this reason, while the exact observed levels of mutations and surviving mutants with specific amino acid changes may have also been influenced by biases in somatic mutation targeting or sequencing error, these explanations could not be the only reasons for our results. It would be unreasonable to think that mutation bias and sequencing error would only influence non-synonymous mutation rates and so it is thus quite clear that selection is causing this skew in mutant numbers. While assessing the exact rate of selection is beyond the scope of this paper, we can attempt to use these levels of synonymous mutations to estimate some ballpark level of expected non-synonymous mutations, which under neutral conditions we would assume to be three times as high. We can then see that all the positions with lowest germline diversity must be undergoing quite stringent negative selection and that once germline diversity gets higher (>2) there are some positions that appear to also be undergoing some positive selection. The

positions with a germline diversity value >5 show rates of non-synonymous mutations 10- to 20-fold greater than synonymous mutations – a clear indicator of strong positive selection.

Another indication that specific positive selection has great influence on the final level of amino acid changed at each position is that the diversity of changed positions is much higher than the germline diversity at those positions (**Figure 3B**). Furthermore, in most cases their bias in usage is indeterminate (**Figure 4**). Thus, while the likelihood of survival is related to germline diversity, the specific change in amino acid that is needed to save the clone is also determined by the specific selection interactions in which that change was positively selected. However, it is worth noting that positions that can be classified (i.e., are not indeterminate) in the mutants all exhibit the same general amino acid bias as the germline repertoire (**Table 1**).

Taking all of these findings into account, we propose that germline diversity is a good indicator of the likelihood of survival following mutation but cannot account for the specific amino acid whose usage accounted for survival of a specific clone, although this usage can be approximated. This usage is based on the specific affinity maturation event and immune response that leads to the formation of the clone. We would further conclude that while CDR and FR do roughly segregate the sequence, a better measure of potential selection force is the specific germline diversity of each position. This is especially true for positions with <5 diversity in their germline amino acids. In such positions, while diversity indicates a range of possible levels of surviving mutants, there is no clear distinction between positions in the CDR and the FR. Indeed, the only reason one exists beyond diversity of 5 is that no FR positions have such high germline diversities.

ACKNOWLEDGMENTS

The authors would like to thank Deborah Dunn Walters for giving us access to her data from her experiments and with several interesting conversations on the meaning of its behavior. Funding: Gregory W. Schwartz is funded by the Graduate Assistance in Areas of National Need (GAANN) program.

REFERENCES

1. Alt FW, Baltimore D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc Natl Acad Sci U S A* (1982) **79**(13):4118–22. doi:10.1073/pnas.79.13.4118
2. Sakano H, Huppi K, Heinrich G, Tonegawa S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* (1979) **280**(5720):288–94. doi:10.1038/280288a0
3. Weigert M, Gatmaitan L, Loh E, Schilling J, Hood L. Rearrangement of genetic information may produce immunoglobulin diversity. *Nature* (1978) **276**(5690):785–90. doi:10.1038/276785a0
4. Kim S, Davis M, Sinn E, Patten P, Hood L. Antibody diversity: somatic hypermutation of rearranged VH genes. *Cell* (1981) **27**(3):573–81. doi:10.1016/0092-8674(81)90399-8
5. Shlomchik MJ, Marshak-Rothstein A, Wolfowicz CB, Rothstein TL, Weigert MG. The role of clonal selection and somatic mutation in autoimmunity. *Nature* (1987) **328**(6133):805–11. doi:10.1038/328805a0
6. Caton AJ, Swartzentruber JR, Kuhl AL, Carding SR, Stark SE. Activation and negative selection of functionally distinct subsets of antibody-secreting cells by influenza hemagglutinin as a viral and a neo-self antigen. *J Exp Med* (1996) **183**(1):13–26. doi:10.1084/jem.183.1.13
7. Casola S, Otipoby KL, Alimzhanov M, Humme S, Uyttersprot N, Kutok JL, et al. B cell receptor signal strength determines b cell fate. *Nat Immunol* (2004) **5**(3):317–27. doi:10.1038/ni1036
8. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) **132**(2):211–50. doi:10.1084/jem.132.2.211
9. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, et al. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol* (2009) **183**(11):7314–25. doi:10.4049/jimmunol.0902452
10. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) **20**(5):683–94. doi:10.1093/intimm/dxn045
11. Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, et al. A Shannon entropy analysis of immunoglobulin and t cell receptor. *Mol Immunol* (1997) **34**(15):1067–82. doi:10.1016/S0161-5890(97)00130-2
12. Schwartz GW, Hershberg U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. *Phys Biol* (2013) **10**(3):1–10. doi:10.1088/1478-3975/10/3/035005
13. Lefranc M-P. IMGTÂ®, the international ImMunoGeneTics information systemÂ® for immunoinformatics. Methods for querying IMGTÂ® databases, tools and webresources in the context of immunoinformatics. *Mol Biotechnol* (2008) **40**(1):101–11. doi:10.1007/s12033-008-9062-7
14. Wu Y, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front Immunol* (2012) **3**:193. doi:10.3389/fimmu.2012.00193
15. Heck LK, van Belle G, Simberloff D. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* (1975) **56**(6):1459–61. doi:10.2307/1934716
16. Hershberg U, Shlomchik MJ. Differences in potential for amino acid change following mutation reveals distinct strategies for kappa and lambda light-chain variation. *Proc Natl Acad Sci U S A* (2006) **103**(43):15963–8. doi:10.1073/pnas.0607581103
17. Hill M. Diversity and evenness: a unifying notation and its consequences. *Ecology* (1973) **54**(2):427–32. doi:10.2307/1934352
18. Jost L. Entropy and diversity. *Oikos* (2006) **113**(2):363–75. doi:10.1111/j.2006.0030-1299.14714.x
19. Chothia C, Gelfand I, Kister A. Structural determinants in the sequences of immunoglobulin variable domain. *J Mol Biol* (1998) **278**(2):457–79. doi:10.1006/jmbi.1998.1653

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 September 2013; paper pending published: 12 September 2013; accepted: 21 October 2013; published online: 08 November 2013.
*Citation: Schwartz GW and Hershberg U (2013) Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. Front. Immunol. **4**:357. doi: 10.3389/fimmu.2013.00357*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Schwartz and Hershberg. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 September 2013
doi: 10.3389/fimmu.2013.00274

Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions

Gilad Liberman[1†], Jennifer Benichou[2†], Lea Tsaban[2], Jacob Glanville[3] and Yoram Louzoun[1,2]*

[1] *Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan, Israel*
[2] *Department of Mathematics, Bar-Ilan University, Ramat Gan, Israel*
[3] *Protein Engineering and Applied Quantitative Genotherapeutics, Rinat-Pfizer Inc., South San Francisco, CA, USA*

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Sho Yamasaki, Kyushu University, Japan
Nikolai Petrovsky, Flinders Medical Centre, Australia

***Correspondence:**
Yoram Louzoun, Department of Mathematics, Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel
e-mail: louzouy@math.biu.ac.il

[†] *Gilad Liberman and Jennifer Benichou have contributed equally to this work.*

Affinity maturation occurs through two selection processes: the choice of appropriate clones (clonal selection), and the internal evolution within clones, induced by somatic hyper-mutations, where high affinity mutants are selected for. When a final population of immunoglobulin sequences is observed, the genetic composition of this population is affected by a combination of these two processes. Different immune induced diseases can result from the failure of regulation of clonal selection or of the regulation of the within clone affinity maturation. In order to understand each of these processes separately, we propose a mixed lineage tree/sequence based method to detect within clone selection as defined by the effect of mutations on the average number of offspring. Specifically, we measure the imbalance in the number of leaves in lineage trees branches following synonymous and non-synonymous (NS) mutations. If a mutation is positively selected, we expect the number of leaves in the sub-tree below this mutation to be larger than in the parallel sub-tree without the mutation. The ratio between the number of leaves in such branches following NS mutations can be used to measure selection within a clone. We apply this method to the sampled Ig repertoire from multiple healthy volunteers and show that within clone selection is positive in the CDR2 region and either positive or negative in the CDR3 and FWR3 regions. Selection occurs already at the IgM isotype level mainly in the DH gene region, with a strong negative selection in the join region. This is followed in the later memory stages in the CDR2 region. We have not studied here the FWR1 and CDR1 regions. An important advantage of this method is that it is very weakly affected by the baseline mutation model or by sampling biases, as are most synonymous to NS mutations ratio based methods.

Keywords: adaptive evolution, phylogenetic tree, immune system, micro-evolution, tree shapes

INTRODUCTION

The humoral adaptive immune response is based on the production of high affinity antibodies against pathogenic antigens. These antibodies are produced through an affinity maturation process, where high affinity antibodies are produced from a large number of cells with different B Cell Receptors (BCRs). The affinity maturation process involves two stages. The first one is clonal selection, where a set of cells with initially mid-high affinity of their BCR to the antigen are expanded. Following this stage, each clone passes a within clone increase in affinity, through somatic hypermutation (SHM) (1) that alter the properties of the BCR, mainly in the complementarity determining region (CDR)3 region (2). The end product of this affinity maturation process is a large population of B cells, with varying BCR that often contain at least one high affinity clone (3).

While the choice of a clone is equivalent to the selection of one specie among many, the dynamics of specific clones in the B cell response against pathogens (4, 5) is a classical example of a process involving rapid asexual reproduction, where constant diversification and adaptation occurs following a high mutation rate. While many tools to study the evolution of species have been developed, tools for the analysis of within population evolution are lacking. We here propose a new method to analyze the within clone affinity maturation process and use it to analyze the healthy B cell repertoire in a large cohort of healthy volunteers.

Multiple methods have been proposed for the measures of selection in populations or between populations, in the sense of evolving toward a higher fitness phenotype (**Table 1**). A now classical measure is the synonymous (S) to non-synonymous (NS) mutations. Specifically, a comparison of the observed and expected NS/(NS + S) ratios is often used as a measure for selection. The expected ratio is calculated based on an underlying mutation probability model [e.g., Ref. (6–8)], or based on genetic regions where no selection is assumed to occur (9). An increased frequency of NS mutations is treated as an indication for positive selection and a decreased one indicates negative selection. Important drawbacks of such methods are: (a) their strong sensitivity to the baseline mutation model (i.e., the expected probability of each mutation type), especially when the mutations rate is position dependent, as happens for example in immunoglobulin sequences (10), and (b) the effect of sampling biases. However, the main problem lies in

Table 1 | List of existing methods based on their reference (first column), the method they use (second column), whether they can detect the direction of selection (third column) and the baseline to which they compare in order to define if selection took place (last column).

Reference	Method	Directional	Baseline reference
Nei and Gojobori (6)	S vs. NS ratio	Yes	Mutation probability model
Yang (7)	S vs. NS ratio	Yes	Mutation probability model
Yang and Nielsen (8)	S vs. NS ratio	Yes	Mutation probability model
Shlomchik et al. (9)	S vs. NS ratio	Yes	Genetic region with no selection pressure
Hershberg et al. (10)	S vs. NS ratio	Yes	Position dependent mutation probability model
Sackin (11)	Tree morphology	No	Yule model
Colless (12)	Tree morphology	No	Yule model
Tajima (16), others	Mutation statistic	No	Naive evolution

the fact that they were developed for an analysis of the comparison between species (clones in this case), and not within a specie.

A different approach for detecting selection is to use properties of lineage trees. Two of the most powerful such methods are Sackin's and Colless's statistics (11–14). Sackin's index is the average root'-leaf distance (over all leaves). Colless's index is the sum of imbalance over all nodes, where a node's imbalance is taken to be the difference in number of leaves between the bigger and smaller sub-trees. These measures are tested vs. a neutral model, which is usually the Yule model, where a tree is constructed by giving each branch the same probability to split (15). Other statistics do not use trees but are based on properties of the full sequences, most notably Tajima's D (16). Such methods have two well-known limitations. They do not distinguish between S and NS mutations and statistical power is lost. However, perhaps the most significant limitation is that in most cases, these methods cannot differ between different types of selection, e.g., positive and negative ones.

We here offer a more direct approach to measure selection within a clone, as well as a better definition of its meaning. This new method overcomes limitations of the S to NS mutation ratio and of the tree shape based selection detection methods, by accounting for the completing information found in each of the two, that is, the classification into mutation types, and the imbalance between different sub-trees.

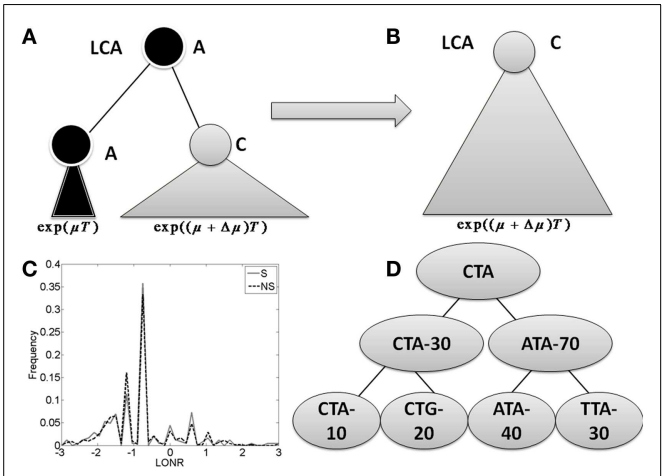


FIGURE 1 | The branch imbalance framework and examples. **(A)** Schematic view of a branch corresponding to a mutation event. Following a mutation, the population can be expanded (or reduced), the advantage will lead to an exponentially growing difference in the number of offspring in parallel branches descending from the same internal origin. *T* is the time from the mutation to the sampling time. **(B)** After some time, one branch will take over the entire sample, and the information carried in the ratio between the branches will be lost. **(C)** LONR values histogram for one simulated sequence pool, simulated under naive multiplication from unique ancestral sequence. While the average is not 0, there is no difference between branches following S and NS mutations. **(D)** Example of tree. In the left branch, a mutation occurred from CTA to CTG, and the ratio between the mutated and un-mutated branches number of offspring is 20/10. In the right branch, a mutation from ATA to TTA occurred, with a ratio of 30/40. In the root, a mutation from CTA to ATA occurred with a ratio of 70/30.

MATERIALS AND METHODS

SELECTION SCORE

Given a tree, each mutation event is assigned: (a) a NS or S mutation flag by its effect on the amino-acid translation of the containing codon; (b) the location of the mutation (related gene where applicable, and number of nucleotides from the beginning of the sequence, otherwise); and (c) the log of the ratio between the number of leaves (sequences) in the sub-tree following the mutation branch and the number of leaves in the sub-tree following the non-mutated branch (see **Figure 1**; **Figure A1** in Appendix). This ratio is denoted the Log Offspring Number Ratio (LONR). This log-ratio is thus positive if the number of final sequences marked by the tree construction algorithm as descendants of the mutated sequence is larger than the number of final sequences marked as descendants of the non-mutated sequence. This suggests some better fitness of the mutated sequence, or positive selection for such mutation, and negative in the opposite case. For each area of the sequence, a *t*-test is performed (unpaired, unequal variances) between the NS and S mutations (see **Figure A4** in Appendix for flow chart).

SIMULATION

A sequence pool simulating neutral reproduction was generated from a random original sequence of 348 nt, with a constant multiplication rate of two offspring per organism. Two regions were defined with uniform mutation probabilities with average mutation rate of 1/2 and 1 mutation per generation. The population was sampled in different sample sizes and along different generations.

In each sampling, one of the first eight siblings (the third generation) was chosen randomly, and its descendants had a twice higher probability of being sampled, effectively simulating sampling bias for a specific clone. The process was repeated 1000 times, and selection was computed in the described process. NS and S mutations were defined relative to their direct ancestor, resulting in unequal NS and S probabilities. All mutations had similar probabilities (i.e., we did not differentiate between Purines and Pyrimidines).

STATISTICAL ANALYSIS

For the immunoglobulin data, the receptors where clustered by isotype (IgA and IgG). Lineage trees were constructed and the sequences were divided into CDR and FWR regions. Mean LONR NS-S difference was computed per clone and per region along with two sample t-test p-values.

IMMUNOGLOBULIN SEQUENCES

Over 500,000 BCRs were sampled from each donor in 12 donors (17), using 454 sequencing, and a RACE protocol. The details of the sequencing and the validity checks are beyond the scope of this manuscript. For each sequence, the most fitting VH, JH, and V-J distance was found by maximizing the relative number of non-mutations for both VH and JH segments. Only sequences that matched higher than 0.5 in both segments were kept for further analysis. The sequences were then clustered according to the most fitting VH and JH as well as the distance between VH and JH, and were truncated to 159 nt from the end of the germline VH and 20 nt from the beginning of the germline JH. Only trees with a sufficient number of mutations were analyzed, which automatically removed sequences defined with ORF and pseudo VH genes (**Figure A3** in Appendix).

CLONE DEFINITION

Sequences were grouped into clones using a two-step approach. First, the germline VH and JH of each sequence were determined by aligning all possible germline VH and JH (based on the IMGT germline library) (18) against the sequence using the Basic Local Alignment Search Tool (BLAST) (19).

Next, in order to count the clones, we grouped all sequences according to their VH and JH usage as well as the distance between VH and JH, since SHMs usually do not produce additions or deletions of nucleotides. Thus, every clone emerging from the same founder cell should have the same distance between VH and JH. We then took all of the sequences with the same VH, JH, and distance between VH and JH and grouped them using a phylogenic approach. The distance between VH and JH was computed by positioning the IMGT germline VH and JH genes on the observed sequence and determining the distance between the last nucleotide of VH and the first nucleotide of JH.

All the sequences with equal VH, JH, and distance were aligned together with an artificial sequence composed of the germline and gaps between them. Within each group, the sequences were aligned (using MUSCLE 3.6) (20), and a phylogenetic tree was built using maximum parsimony (21) and/or neighbor joining (22) methods (from the PHYLIP 3.69 program package). We then parsed this tree with a cutoff distance of four mutations into clones. Thus, a clone was defined as a set of sequences that are similar one to each other, up to a distance of four mutations.

RESULTS

SELECTION

Before discussing B cells specifically, let us discuss how selection can be estimated in a rapidly mutating population. Assume a population originating from a single founder through asexual division. In the case of B cells, this would be a clone seeded by an ancestral B cell with a given H chain rearrangement. We ignore the L chain at this stage. The genetic sequence of the founder can be changed by mutations that can affect the population dynamics. In such a case we would define positive selection in the population as an increased average division/birth rate or a decreased average death rate following mutations. Note that these are not precisely the same, especially in the context of B cell dynamics (23), but this is beyond the scope of the current analysis. A decrease in the division rate would be defined as negative selection. Note that each mutation by itself can have a positive, null, or negative effect, but the definition of selection is based on the average population dynamics and not with the dynamics following a single mutation.

Let us follow a mutation that occurs within a population. If this mutation increases the average number of offspring per generation from μ to $\mu + \Delta\mu$, then by a time proportional to log of the total population size, the advantageous mutation will take over the population. When we compare the population to its latest common ancestor (LCA), we will have no evidence that such a mutation has occurred (**Figures 1A,B**). If the original sequence cannot be known (as often occurs in the CDR3 region), we will not be able to detect the presence of such a mutation. If the original sequence is known (as typically occurs in mutations within the germline VH gene), the genetic composition of the population would be equivalent to the one expected in a neutral model (model with no selection). The only difference would be the addition of a single NS mutation to a gene in the entire population. This information can be used to infer that selection has taken place. This is basically the logic behind S to NS mutation ratio tests for selection.

During the intermediate period when the two sub-clones still exist (the mutated and the un-mutated one), one can compare the population size of the two sub-clones. We expect the ratio between the two population sizes to be proportional to $e^{\Delta\mu T}$, where T is the time from the mutation to the sampling time (**Figure 1A**).

For a single mutation, it will be hard to differentiate between the effect of selection and a non-uniform sampling where one branch is sampled more deeply. However, if many mutations occur in the genetic region of interest, and if on average mutations in this region increase the average number of offspring, we expect more offspring in branches that follow a mutation in this region than in branches emerging from the same direct ancestor with no mutations, and inversely in the case of negative selection.

We thus propose to detect selection using this imbalance in cases where the total mutation rate (mutation rate per organism multiplied by population sample size) is significantly higher than one, as typically occur in within clone B cell evolution.

LOG OFFSPRING NUMBER RATIO

We define a measure of selection in a gene as the ratio of the number of leaves (measured descendants) under the branch where a mutation occurred and the number of decedents in its direct sibling where no such mutation occurred. We compare the

distribution of these ratios (more precisely the log of the ratios) in S and NS mutations to estimate whether the distribution deviates from the one induced by neutral drift (**Figure 1D**).

Specifically, for each mutation occurring in one son of an internal node and not in the other, we compute the sub-tree size under the son with a mutation and the sub-tree under the son without a mutation. The log of the ratio between these two sizes is defined as the LONR of this mutation. We then compute the LONR value for all S and NS mutations in the tree, and compare the S and NS LONR distributions (**Figure 1C**; **Figures A1** and **A2** in Appendix).

Note that this analysis is not sensitive to the details of the baseline model for the probability of either silent or replacement mutations, since their absolute number is never used in the analysis. The only case where such a model would affect the current measurement is in the extreme case that the probability for S would differ by orders of magnitude from the probability of NS mutations.

In order to check that the LONR does not detect selection in its absence, we simulated mutating clones, sampled the resulting sequences (see Materials and Methods for details), produced lineage trees, and compared the LONR distribution following S and NS mutations. When the number of mutations is very small, or the number of samples is small, the False Positive (FP) rate (the cases where the LONR average is significantly different following S and NS mutations with a p-value of 0.05) is higher than the expected 5%. However, in the regime of over 10–20 mutations per sequence and at least 300 sequences per tree, the FP rates are near the expected 5% (data not shown). We have repeated the analysis with non-uniform mutation rates (position dependent mutation rates) and with sampling biases, and obtained similar results, as long as the S and NS mutation rates are of the same order of magnitude. A detailed methodological analysis of the LONR will be given in a separate analysis.

SELECTION ALONG HUMAN Ig CLONES

We have used the LONR score to analyze the healthy repertoire from 12 donors. In such a repertoire two opposite forces operate: (a) mutations can ruin the functionality of the receptor and decrease its survival probability, (b) mutations can on the other hand increase the affinity to the antigen and thus lead to a higher division rate. The CDRs of the BCR determine its interaction with the antigen, and mutations there were reported to have a higher probability to increase the affinity than mutations in the framework (FWR) region (5, 24). However, the net selection effect in each of these regions still remains unclear. Beyond the effect of SHM, B cells are affected by isotype switches from naïve IgM to memory IgM, and from there to memory IgG and IgA. The memory (IgM, IgG, and IgA) isotypes occur at the advanced stages of the immune response and thus lineage trees based on such receptors are expected to represent the full evolution following selection.

We have used high-throughput sequencing to sequence over 500,000 BCR samples from each donor, in 12 donors. We built lineage trees from the sequences [see (17) for details of sequences, and production of lineage trees]. We measured the LONR distribution in all naïve and memory, IgM as well as IgA and IgG sequences trees (over 50,000 lineage trees) and compared the LONR distribution in NS and S mutations. The results are actually quite striking. We analyzed separately CDR2, FWR3, and CDR3, using the standard IMGT definition (18).

We did not analyze the CDR1 and FWR1 and FWR2 regions, since we did not have enough samples with reliable sequences in these regions. Thus, our results only apply to the comparison of the more 3′ regions (CD2,3 and FWR3). Also, we here include the JH region within the CDR3 analysis. This was done in order to avoid artifacts of the DH gene length. In this specific point, our notation slightly deviates from the standard IMGT analysis that ends the CDR3 region at the beginning of the JH gene region.

As expected in both IgG and IgA memory cells, the positive selection is much stronger for the CDR regions than for the FWR (**Figure 2**). However, in the memory IgM, even the FWR region passes a positive selection during the immune response. Thus, one cannot conclude as a generic conclusion that FWR regions are under negative selection, while CDR regions are under a positive selection. Within different CDR regions, CDR2 is under a much more stringent positive selection than the CDR3 region. Highlighting the fact that while the CDR3 is selected at clone level, where clones with an appropriate CDR3 sequence are selected for expansion, mutations in the CDR2 region may induce a much stronger positive selection than mutations in the CDR3 that can induce a more balanced effect. When analyzing only trees where there is a significant difference between the LONR score ($p < 0.01$) of S and NS mutations, the results are qualitatively similar to the analysis of all trees (**Figure 2**, inset).

In order to ensure that the LONR is not an artifact of the sampling depth or the number of mutations, we computed a correlation between the average S vs. NS mutations LONR difference in each sample (a sample being defined as a single donor, a given isotype, and a given VH gene) and the sample size, or the average number of mutations in each region separately or in all regions in this sample. In all cases there was practically no correlation (the

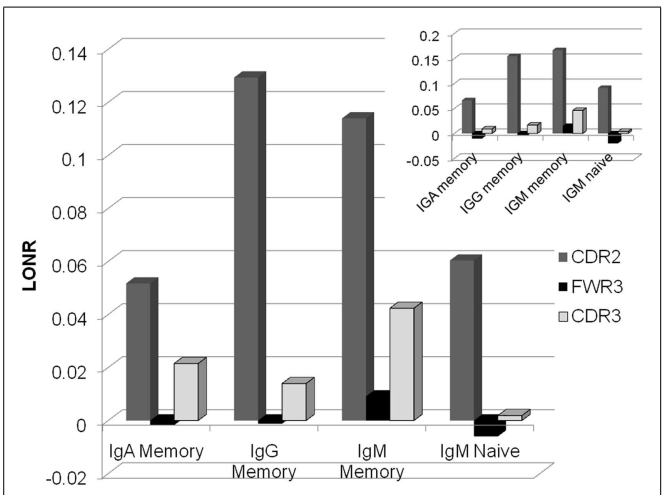


FIGURE 2 | Average LONR score per region and per isotype for all lineage trees (main figure) and for trees with a significant difference (*t*-test, $p < 0.01$; inset). The main positive selection occurs in CDR2, followed by CDR3. FWR3 has a limited if any negative selection. Selection occurs mainly in the memory isotype, but some systematic selection is already observed in the IgM naïve isotype.

highest Spearman correlation was $R = 0.05$). The same occurs if S or NS mutations are used separately. Thus, the observed selection effect is not a sampling or a mutation rate effect.

COMPARISON OF ISOTYPES

Selection and mutations are accompanied by an isotype switch process. It is not clear from the current literature whether isotype switch precedes mutations and selection or if it occurs in parallel. If selection occurs only following isotype switch, we expect no selection to be observed in lineage trees composed of purely IgM sequences. However, the selection level in IgM memory trees is as high as the one observed in IgG lineage trees, and higher than the one observed in IgA lineage trees. Moreover, even in trees composed only of naïve (CD27−) sequences (25, 26), a clear advantage in the division rate following mutations in the CDR2 region is observed. One can thus conclude that even before they become memory cells, B cells pass an antigen induced selection. Note that the CD27− cells can be activated cells, and are not of a pure naïve type (27, 28).Thus, the observed mutations and selection may actually represent an activated phenotype which is a part of the CD27− sub population.

COMPARISON BETWEEN DONORS AND BETWEEN VH GENES

The results presented in **Figure 2** are the average of the selection score over many donors. Some variability exists between donors, and the average selection score can represent a combination of positive selection in some donors and negative selection in others. We have thus separated the analysis into different donors (**Figure 3**). A clear difference emerges between the different regions. While in the CDR2 region practically all donors show a positive selection in all isotypes, in FWR3 and CDR3, the results are highly variables in all isotypes, with some donors showing a marked positive selection and some a negative selection. The main difference between FWR3 and CDR3 is not in the sign of the selection but in its variance. In CDR3 the variance among donors is much larger than in FWR3.

POSITION EFFECT

A more complex picture emerges when each position is analyzed by itself, instead of merging all positions belonging to the same region. At the naïve IgM level, positive selection is mainly focused on the DH region within the CDR3 region, while negative selection takes place in the junction regions (**Figure 4**). When moving to the IgM memory isotype the selection in the CDR3 becomes much weaker, and the selection in the CDR2 starts to rise. Finally, when moving to the IgA and IgG isotype selection is fully focused on CDR2. Note that we here plot the net selection per nucleotide, without considering the total number of mutations per nucleotide. Thus some nucleotide may contribute significantly more than other nucleotides to the average. This different weighting induces some quantitative differences between **Figures 3** and **4**. Note also that while the total number of mutations per sequences is much larger in IgG and IgA than in the naïve IgM serotype, the selection is actually maximal at the naïve level.

These results are highly consistent among the different independent donors, with selection patterns practically overlapping in 10 donors out of 12 (**Figure 4**). In two donors the observed

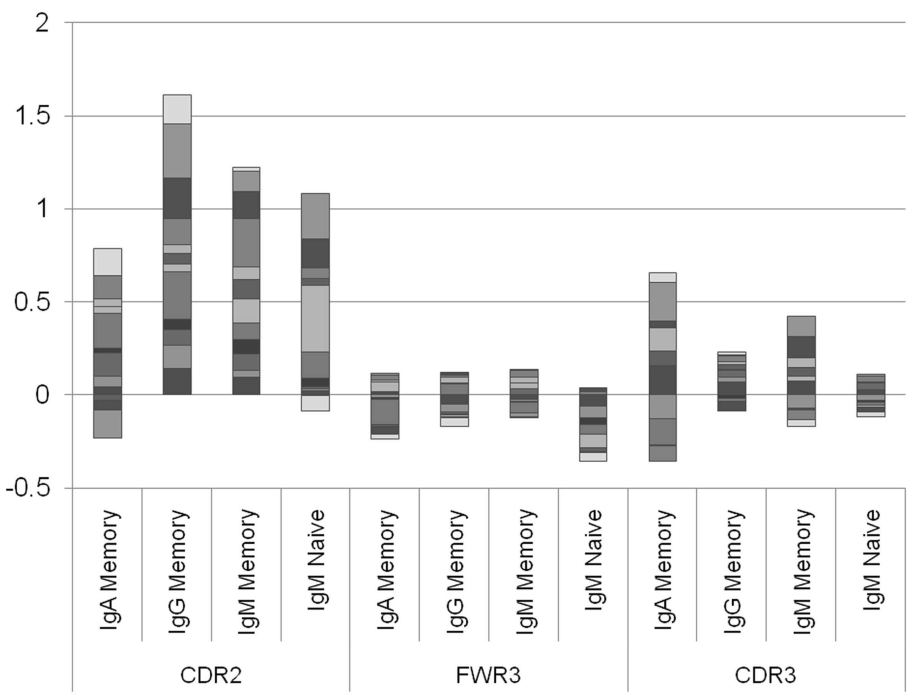


FIGURE 3 | Log offspring number ratio per donor and per isotype. Each column is an isotype in a different region, and each color is a different donor. Each column represents the aggregate over many donors, where negative and positive values are drawn separately. While in the CDR2 there are practically no donors where the average selection is negative (with the exception of two IgA samples and one naïve IgM sample), in the CDR3 there are approximately the same number of donors with negative selection and with positive selection. This shows that the selection in CDR2 is a universal feature, while in CDR3 it may depend on the random junction initially produced or in the different exposures to antigens.

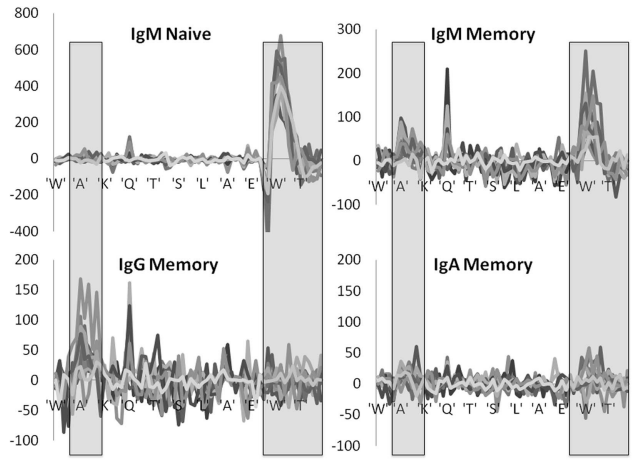


FIGURE 4 | Position specific effects. Total LONR score per amino acid (averaged over the 3-nt composing the amino acid). The LONR is drawn for the four isotypes discussed in previous figures: naïve IgM, memory IgM, IgG, and IgA. Each line represents a different donor. The highlighted regions represent the CDR2 (to the left) and CDR3 (to the right). The sequence at the bottom is a typical sequence. The zone of very strong negative selection at the beginning of the CDR3 is the junction region. One can observe a switch from a very strong positive selection in the naïve IgM isotype focused on the CDR3 to a much weaker selection in the IgG and IgA focused on the CDR2. Note that the scales of the y axes are different between the plots.

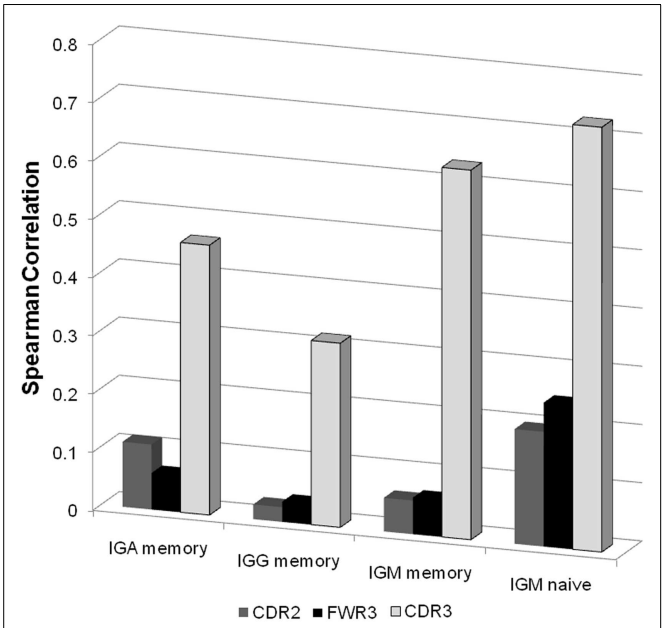


FIGURE 5 | Spearman correlation between the NS/(S + NS) fraction and the total number of B cells for different isotype and different regions. For each sample we produced a vector of 46 values (all functional VH genes with enough samples) representing the total number of B cells in this sample with such a V gene. When all samples are taken together, this leads to a 12*2*46 values (12 donors, 2 technical repeats per donor). We computed the correlation of these values with the NS/(S + NS) values in the same categories. As one can clearly see the highest correlation is with the CDR3 regions of naïve IgM samples.

selection was weak and noisy, but these donors had much smaller sample sizes (data not shown).

These results suggest the following two stage selection process occurring in Ig lineage trees. The first stage of selection occurs early in the CDR3 and is generic and uniform. This would be equivalent to the "key mutation" concept (29, 30). Following these key mutations, selection becomes much weaker and focused on the CDR2 regions in the IgG isotype. The memory IgM shows a translational mutation distribution from the key-mutation selection event, to the weaker mutation in the CDR2, which may alters the affinity in a limited way.

In order to check that there is indeed a correlation between mutations in the CDR3 regions of naïve cells, we compared the correlation between the NS/(S + NS) ratio and the total number of B cells for different regions, VH genes, and isotypes (**Figure 5**). The total clone size was defined as the number of sequences with a give VH in a give sample. The NS and S mutations were defined as the average number of mutations in the leaves compared with the ancestral sequence of each clone (a similar analysis with the number of unique mutations led to similar results). The total number of B cells is the total number of B cells sequenced in a given sample with the same VH gene. As expected the highest correlation was with the NS/(S + NS) ratio in the CDR3 region of naïve IgM cells. Note that a trivial correlation is expected between the number of B cells and the total number of mutations, since large clones may be older clones and as such accumulate a large number of mutations. However, here we have compares the NS fraction, which is not expected to be correlated with the population size, unless selection is involved.

DISCUSSION

While multiple sequence based methods have been proposed to detect selection (10, 31–36), most of them are sensitive to the baseline mutation model, or to sampling effects. Moreover, many existing methods conclude the existence of positive or negative selection from highly frequent or rare mutations. Such an under or over expression of a sequence can simply represent the random expansion of a population, a bottleneck effect, or the random association of this mutation with another mutation which is selected. Indeed even in models of neutral evolution, alleles carrying some sequences are expected to be much more frequent than others, since alleles, and sequence distributions have a fat tail. Instead, positive selection should be defined by the systematic increase of the population size following replacement mutations in a given region.

We have here used this precise definition to define the LONR score as the increase/decrease in the relative branch size following a mutation. A comparison between the LONR score following synonymous and NS mutations. A detailed methodological analysis is left to a different framework; we here describe an application of the LONR score to immunoglobulin clones in healthy volunteers.

The LONR differs in two basic aspects from most other S to NS mutation frequencies methods. First this method does not count the absolute number of sequences; instead it measures for each mutation the effect it has on its number of offspring. The second difference is the definition of mutations and their classification as

S/NS in respect to their direct ancestor, and not to a consensus or ancestral sequence.

Since each mutation is counted once, independent of the total number of sequences that end up containing this mutation, it is practically not affected by sampling biases or by the expansion of specific sub-populations. Moreover the LONR does not require a baseline mutation model, since it does not compare the number of synonymous or NS mutations. Instead, it measures the effect of each mutation on the total number of offsprings.

Tree shape based methods were developed (11–14). However, these methods often cannot detect the direction of selection, and cannot detect which region in the sequence is selected. Moreover, many of these tree shapes are sensitive to sampling effects making them impractical to use in realistic situations (37).

The observed selection in Ig clones has a well conserved pattern among donors. It is consistently positive in the CDR2 region, and positive in average in the CDR3 region. The mutation pattern in the CDR3 region is composed of strong positive selection in the DH region, and strong negative selection in the junctions between the VH and DH genes and between DH and JH genes. We currently have no clear explanation for the negative selection in the junction. However, one can hypothesize that only B cells with appropriate junctions are selected to pass affinity maturation, and that following within clone selection must maintain these junctions. Such a behavior has been clearly observed in H chain transgenic models of selection in mice (23). At later stages, selection is focused on the CDR2 region and is much lower in the CDR3 regions. This can represent a fine tuning of the affinity, where the main limiting step is the accumulation of key mutations in the CDR3, which is then followed by the expression of more specific mutations in the CDR2 region.

We did not analyze the CDR1 and FWR1 and FWR2 regions, since we did not have enough samples with reliable sequences in these regions. We can only guess that CDR1 should behave approximately like CDR2. FWR1 and FWR2 may be quite different than FWR3 as previously proposed (38). Advances in B cell sequencing technologies (39) will hopefully provide longer reads allowing us to study these regions as well.

REFERENCES

1. Nossal G. The molecular and cellular basis of affinity maturation in the antibody response. *Cell* (1992) **68**:1. doi:10.1016/0092-8674(92)90198-L
2. Kocks C, Rajewsky K. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. *Proc Natl Acad Sci U S A* (1988) **85**:8206–10. doi:10.1073/pnas.85.21.8206
3. Berek C, Milstein C. Mutation drift and repertoire shift in the maturation of the immune response. *Immunol Rev* (1987) **96**:23–41. doi:10.1111/j.1600-065X.1987.tb00507.x
4. Liu Y, Joshua D, Williams G, Smith C, Gordon J, Maclennan I. Mechanism of antigen-driven selection in germinal centres. *Nature* (1989) **342**(6252):929–31. doi:10.1038/342929a0
5. Berek C, Berger A, Apel M. Maturation of the immune response in germinal centers. *Cell* (1991) **67**:1121–9. doi:10.1016/0092-8674(91)90289-B
6. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* (1986) **3**:418–26.
7. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* (1998) **15**:568–73. doi:10.1093/oxfordjournals.molbev.a025957
8. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* (2000) **17**:32–43. doi:10.1093/oxfordjournals.molbev.a026236
9. Shlomchik MJ, Aucoin AH, Pisetsky DS, Weigert MG. Structure and function of anti-DNA autoantibodies derived from a single autoimmune mouse. *Proc Natl Acad Sci U S A* (1987) **84**:9150–4. doi:10.1073/pnas.84.24.9150
10. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) **20**:683–94. doi:10.1093/intimm/dxn026
11. Sackin M. "Good" and "bad" phenograms. *Syst Biol* (1972) **21**:225–6. doi:10.1093/sysbio/21.2.225
12. Colless D. Phylogenetics: the theory and practice of phylogenetic systematics. *Syst Zool* (1982) **31**:100–4. doi:10.2307/2413420
13. Kirkpatrick M, Slatkin M. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* (1993) **47**:1171–81. doi:10.2307/2409983
14. Blum MGB, François O. On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Math Biosci* (2005) **195**:141–53. doi:10.1016/j.mbs.2005.03.003
15. Yule GU. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Phil Trans R Soc Lond B* (1925) **213**:21–87.
16. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* (1989) **123**:585–95.
17. Benichou J, Glanville J, Prak ETL, Azran R, Kuo TC, Pons J, et al. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol* (2013) **190**:5567–77. doi:10.4049/jimmunol.1201929
18. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* (1999) **27**:209–12. doi:10.1093/nar/27.1.209
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* (1990) **215**:403–10. doi:10.1016/S0022-2836(05)80360-2
20. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* (2004) **32**:1792–7. doi:10.1093/nar/gkh340
21. Kolaczkowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* (2004) **431**:980–4. doi:10.1038/nature02917
22. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* (1987) **4**:406–25.
23. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, et al. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol* (2009) **183**:7314–25. doi:10.4049/jimmunol.0902452
24. Cowell LG, Kim HJ, Humaljoki T, Berek C, Kepler TB. Enhanced evolvability in immunoglobulin V genes under somatic hypermutation. *J Mol Evol* (1999) **49**:23–6. doi:10.1007/PL00006530
25. Klein U, Rajewsky K, Küppers R. Human immunoglobulin (Ig) M+ IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J Exp Med* (1998) **188**:1679–89. doi:10.1084/jem.188.9.1679
26. Agematsu K, Hokibara S, Nagumo H, Komiyama A. CD27: a memory B-cell marker. *Immunol Today* (2000) **21**:204–6. doi:10.1016/S0167-5699(00)01605-4
27. Fecteau JF, Côté G, Néron S. A new memory CD27-IgG+ B cell population in peripheral blood expressing VH genes with low frequency of somatic mutation. *J Immunol* (2006) **177**:3728–36.
28. Cagigi A, Du L, Dang LVP, Grutzmeier S, Atlas A, Chiodi F, et al. CD27(-) B-cells produce class switched and somatically hypermutated antibodies during chronic HIV-1 infection. *PLoS One* (2009) **4**:e5427. doi:10.1371/journal.pone.0005427
29. Radmacher MD, Kelsoe G, Kepler TB. Predicted and inferred waiting times for key mutations in the germinal centre reaction: evidence for stochasticity in selection. *Immunol Cell Biol* (1998) **76**:373–81. doi:10.1046/j.1440-1711.1998.00753.x
30. Kleinstein SH, Singh JP. Toward quantitative simulation of germinal center dynamics: biological and modeling insights from experimental validation. *J Theor Biol* (2001) **211**:253–75. doi:10.1006/jtbi.2001.2344

31. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* (2000) **15**:496–503.
32. Plotkin JB, Dushoff J, Fraser HB. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature* (2004) **428**:942–5. doi:10.1038/nature02458
33. Wong WSW, Nielsen R. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* (2004) **167**:949–58. doi:10.1534/genetics.102.010959
34. Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. *Genetics* (2005) **169**:1753–62. doi:10.1534/genetics.104.032144
35. Pond SLK, Frost SDW. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* (2005) **22**:478–85. doi:10.1093/molbev/msi031
36. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* (2005) **22**:2472–9. doi:10.1093/molbev/msi237
37. Stam E. Does imbalance in phylogenies reflect only bias? *Evolution* (2002) **56**:1292–5. doi:10.1554/0014-3820(2002)056[1292:DIIPRO]2.0.CO;2
38. Shapiro GS, Aviszus K, Ikle D, Wysocki LJ. Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition. *J Immunol* (1999) **163**:259–68.
39. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* (2012) **135**(3):183–91. doi:10.1111/j.1365-2567.2011.03527.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 June 2013; paper pending published: 16 July 2013; accepted: 28 August 2013; published online: 17 September 2013.

*Citation: Liberman G, Benichou J, Tsaban L, Glanville J and Louzoun Y (2013) Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions. Front. Immunol. **4**:274. doi: 10.3389/fimmu.2013.00274*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Liberman, Benichou, Tsaban, Glanville and Louzoun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

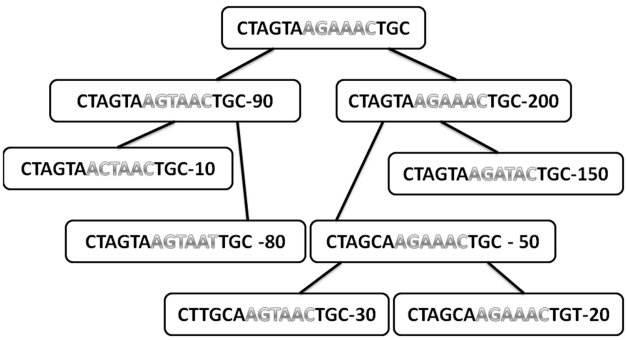


FIGURE A1 | Typical tree with the sequences in each node and the total number of offspring of each node. Only internal nodes are drawn here (leaves are not used for the analysis, since they have no mutations under them). The sequences are divided into two regions (different colors). The number in each node is the total number of offspring.

Position	Mutation	Translation	Type	Zone	LONR
9	AGA->AGT	R->S	NS	2	Log(90/200)
8	AGT->ACT	S->T	NS	2	Log(10/80)
12	AAC->AAT	N->N	S	2	Log(80/10)
10	AAC->TAC	N->Y	NS	2	Log(150/50)
3	CTA->CTT	L->L	S	1	Log(30/20)
15	TGC->TGT	C->C	S	1	Log(20/30)
5	GTA->GCA	V->A	NS	1	Log(50/150)

FIGURE A2 | Log offspring number ratio values for S and NS mutations as given in Figure A1. For each mutation its region is give (1 or 2), its type (S or NS), its position along the gene in nucleotides, the mutation itself in nucleotides, and amino acids, as well as the LONR score. The distribution of LONR scores is used to assess selection.

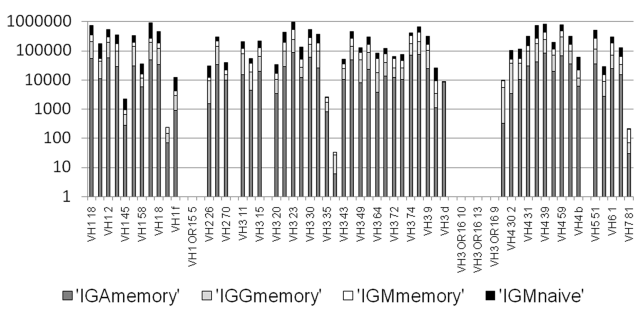


FIGURE A3 | Number of mutation events per VH gene per isotype. There are practically no mutations assigned to pseudogenes, since practically none of the sequenced in-frame lineage trees are based on pseudogenes (<0.05%). However, there are also some functional VH genes with practically no lineage trees. We removed those from the current analysis.

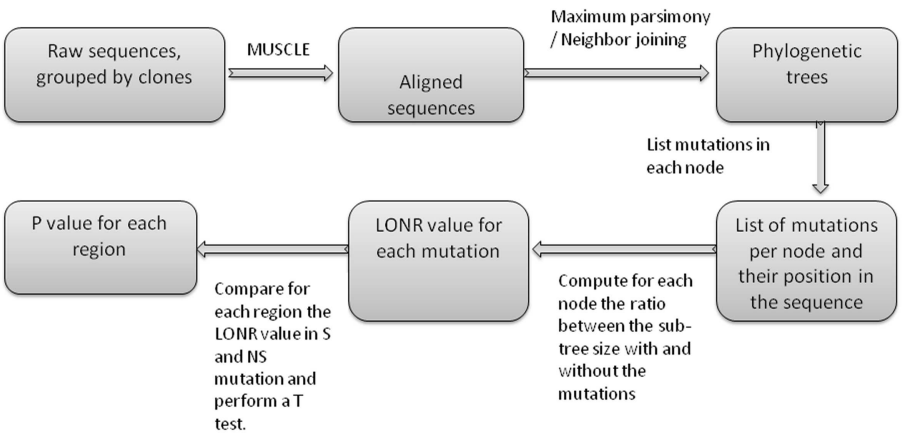


FIGURE A4 | Flow chart of LONR score production. First sequences are aligned, then phylogenetic trees are produced. For each internal node in the trees, a list of mutations is produced (between the internal node and one of its direct descendents). For each such mutation the LONR score is computed resulting in a list of LONR score for each mutation. For genetic region, the LONR score of the mutations in this regions are analyzed and the resulting *p*-value is the result of a *t*-test between S and NS mutations.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 22 April 2014
doi: 10.3389/fimmu.2014.00170

Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation

Thomas B. Kepler[1,2]*, Supriya Munshaw[3], Kevin Wiehe[4], Ruijun Zhang[4], Jae-Sung Yu[4,5], Christopher W. Woods[5,6,7,8], Thomas N. Denny[4,5], Georgia D. Tomaras[4,5,9], S. Munir Alam[4,5], M. Anthony Moody[3,4], Garnett Kelsoe[4,9], Hua-Xin Liao[4,5] and Barton F. Haynes[4,5]

[1] Department of Microbiology, Boston University School of Medicine, Boston, MA, USA
[2] Department of Mathematics and Statistics, Boston University, Boston, MA, USA
[3] Center for Viral Hepatitis Research, Johns Hopkins University, Baltimore, MD, USA
[4] Duke Human Vaccine Institute, Duke University Medical Center, Durham, NC, USA
[5] Department of Medicine, Duke University Medical Center, Durham, NC, USA
[6] Department of Pathology, Duke University Medical Center, Durham, NC, USA
[7] Hubert-Yeargan Center for Global Health, Duke University Medical Center, Durham, NC, USA
[8] Department of Pediatrics, Duke University Medical Center, Durham, NC, USA
[9] Department of Immunology, Duke University Medical Center, Durham, NC, USA

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Andrew M. Collins, University of New South Wales, Australia
Uri Hershberg, Drexel University, USA

***Correspondence:**
Thomas B. Kepler, Departments of Microbiology, Mathematics and Statistics, Boston University, 72 East Concord Street, L-504, Boston, MA 02118, USA
e-mail: tbkepler@bu.edu

Affinity maturation of the antibody response is a fundamental process in adaptive immunity during which B-cells activated by infection or vaccination undergo rapid proliferation accompanied by the acquisition of point mutations in their rearranged immunoglobulin (Ig) genes and selection for increased affinity for the eliciting antigen. The rate of somatic hypermutation at any position within an Ig gene is known to depend strongly on the local DNA sequence, and Ig genes have region-specific codon biases that influence the local mutation rate within the gene resulting in increased differential mutability in the regions that encode the antigen-binding domains. We have isolated a set of clonally related natural Ig heavy chain–light chain pairs from an experimentally infected influenza patient, inferred the unmutated ancestral rearrangements and the maturation intermediates, and synthesized all the antibodies using recombinant methods. The lineage exhibits a remarkably uniform rate of improvement of the effective affinity to influenza hemagglutinin (HA) over evolutionary time, increasing 1000-fold overall from the unmutated ancestor to the best of the observed antibodies. Furthermore, analysis of selection reveals that selection and mutation bias were concordant even at the level of maturation to a single antigen. Substantial improvement in affinity to HA occurred along mutationally preferred paths in sequence space and was thus strongly facilitated by the underlying local codon biases.

Keywords: somatic hypermutation, experimental influenza infection, antibody selection, antibody affinity maturation, phylogenetics

INTRODUCTION

B-cells that respond to infection or vaccination are induced by signaling through their B-cell receptors to proliferate and differentiate into plasmacytes and memory cells. Short-lived plasmacytes secrete antibody and provide immediate protection from the eliciting agent; memory cells and long-lived plasmacytes persist clonally for very long times, providing protection against recurring challenges from the same or closely related agents (1). Cells that go on to find persistent clones are subject to affinity maturation in their post-exposure development. During affinity maturation, the affinity of the B-cell receptor for antigens on the eliciting agent is substantially increased, resulting in a more potent response on recall (2).

Affinity maturation proceeds through somatic hypermutation, the introduction of point mutations into the rearranged immunoglobulin (Ig) genes that encode the B-cell receptor. Those B-cells that thereby acquire an increased affinity for the antigen gain a proliferative advantage and come to dominate the activated B-cell population. Affinity maturation is crucial for humoral immune protection, conferring greater neutralization capacity (3) and opsonization efficiency (4), and is generally correlated with higher vaccine efficacy (5). In fact, lack of effective affinity maturation has been directly implicated in adverse outcomes for at least one vaccine (6).

The rate of somatic hypermutation at a given position with an Ig variable region is significantly influenced by the local DNA sequence – both the nucleotide at that position and sequence of nucleotides containing it (7). Codon usage in Ig V-gene segments is strongly biased, with zones of high mutability largely overlapping with the complementarity-determining regions (CDR), which encode the antibody's antigen-binding residues (8, 9). Thus, somatic mutation drives Ig genes along statistically favored paths through the genotype space. Some combinations of substitutions will therefore be visited much more rapidly than others involving the same number of changes. Each Ig gene segment has been involved in the response to a huge number of antigens over the course of its evolutionary history and has experienced selection pressure to enhance its role as a template for affinity maturation.

Technology for the isolation of native heavy-chain/light-chain pairs and their subsequent recombinant synthesis have recently been developed (10, 11) and refined (12), making it feasible to determine the biophysical properties of large numbers of monoclonal antibodies (mAb). We have complemented this technology with the development of computational tools that substantially improve our ability to infer the unmutated common ancestor of a set of clonally related antibodies, and the corresponding maturation intermediates (13).

We have now applied these methods to the detailed study of the maturation pathways of a B-cell clone whose antibody genes were isolated from a human experimental influenza infection study, providing an elucidation of the interplay of mutational constraints and selection on antigen-binding affinity. One of our aims in this study is to examine the influence that this differential mutability has on a specific instance of affinity maturation to a given antigen: the immune response to influenza hemagglutinin (HA) in a human subject. This question clearly goes beyond the issue of codon bias as a statistical regularity to inquire about influence of codon bias in a specific case. The relationship between these two questions is analogous to the phenomenon of HCDR3 length in autoimmune disease. There one has the statistical observation that B-cells with long HCDR3 are counter-selected during development (14), yet the role of long HCDR3 for individual autoantibodies is rarely understood. In our case, we know that the mutation frequency is higher on average in regions that encode amino acids that are more likely, on average, to contact epitopes. In this study, we examine the interplay of differential mutation frequency and selection in the evolution of a single antibody lineage.

Specifically, we demonstrate that intraclonal affinity maturation proceeded by stepwise accumulation of affinity-enhancing mutations and that mutation and selection interacted synergistically. These insights and others gained by application of the tools we have developed promise to facilitate the effective harnessing of affinity maturation for vaccine engineering.

RESULTS

ISOLATION AND IDENTIFICATION OF ANTI-INFLUENZA HEMAGGLUTININ A B-CELL CLONE CL2569

Human subjects were experimentally infected intranasally with influenza virus (15). Eighty-six natural heavy-chain/light-chain gene pairs were isolated from one subject (subject EI13) on day 4 after exposure. Among these, we found three clonally related sets. Two of the clones contained two antibodies each; the other contained five. The members of this five-member clone, designated CL2569, all bind HA in the $K_d = 1$–20 nM range. Four of these antibodies are of the IgM isotype while the other is IgA1. The light chain in each antibody is Ig kappa. The remainder of this study describes our analysis of CL2569.

The antibodies are highly diversified. The heavy chains have a mean (±SD) pairwise difference of 28.0 ± 5.4 nucleotides (nt) and 16.7 ± 3.6 amino acids (aa); the light chains have an average pairwise difference of 18.0 ± 2.5 nt and 8.2 ± 1.5 aa.

We inferred the unmutated ancestor (UA) and intermediates along the affinity maturation pathways by computing the Bayesian posterior probability mass function on nucleotide states at each position of the heavy and light chains conditioned by the data and the maximum-likelihood phylogram as described in the companion study (13). The mutations acquired along each branch were enumerated and classified according to the IMGT classification (16) (**Table 1**). The UA and all intermediates for both heavy and light chains were synthesized using the same recombinant technology used to synthesize the observed antibodies.

The probable error profile for the heavy-chain UA is shown in **Figure 1**. Briefly, the sum of the probable errors over all positions is 3.2. There are five nucleotide positions where the marginal posterior probability of the modal nucleotide is <0.8, all of which occur in CDR3. Importantly, at these somewhat lower-confidence positions, the inferred modal CDR3 is identical to all five observed sequences. The summed probable errors for each of the inferred intermediates is less than that of the inferred UA and decreases as one gets closer to the observed sequences. The kappa chain UA is known with high confidence. The sum of the probable errors is 0.33.

THE DISSOCIATION CONSTANT DECREASES EXPONENTIALLY WITH UNIFORM RATE OVER THE DURATION OF THE PROCESS

The dissociation constant K_d for binding to HA of the Brisbane strain of influenza virus was measured using ELISA on solutions of monoclonal antibody prepared at known concentrations. K_d

Table 1 | Classification of mutations in CL2569 heavy- and light-chain histories.

	Heavy chains		Light chains	
	Non	Synon	Non	Synon
FR	40	20	14	7
CDR	17	11	11	4

FR, framework region; CDR, complementarity-determining region; Non, non-synonymous; Synon, synonymous.

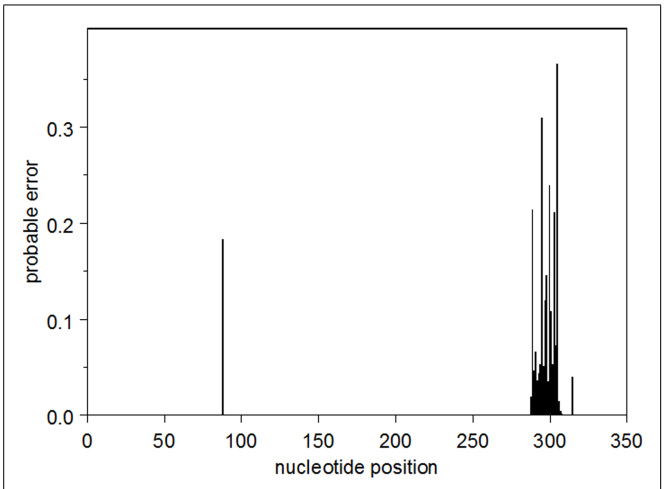


FIGURE 1 | The profile of the probable error in the modal heavy chain UA.

was estimated by non-linear curve fitting simultaneously on all data for each plate. The UA binds to HA very weakly but measurably, $K_d = 2.6\,\mu M$. Throughout the evolutionary process, K_d declines uniformly and exponentially ($R^2 = 0.92$), falling 50–74% (95% confidence interval) for each 1% increase in evolutionary distance (**Figure 2**). The affinities of the observed antibodies are approximately three orders of magnitude higher than that of the ancestor, an improvement that occurs over a total evolutionary distance of 6–9% nucleotide differences.

INTERACTION BETWEEN SELECTION AND MUTABILITY

To gage the force of selection in molecular evolution, deviations in the ratio of the number of synonymous mutations to the number of non-synonymous mutations from that expected under the null hypothesis of selection-free evolution are often used for statistical testing (17). For antibody somatic evolution, mutations are further classified by region, occurring in the CDR or framework regions (FR) and various combinations specific deviations from expected values within these classifications used in statistical tests [see, e.g., Ref. (18)]. Crucially, the distribution expected under the no selection null hypothesis for Ig somatic evolution is not trivially computed. Because the codon bias has been adapted for Ig plasticity, empirical estimation of the distributions under the null cannot be avoided.

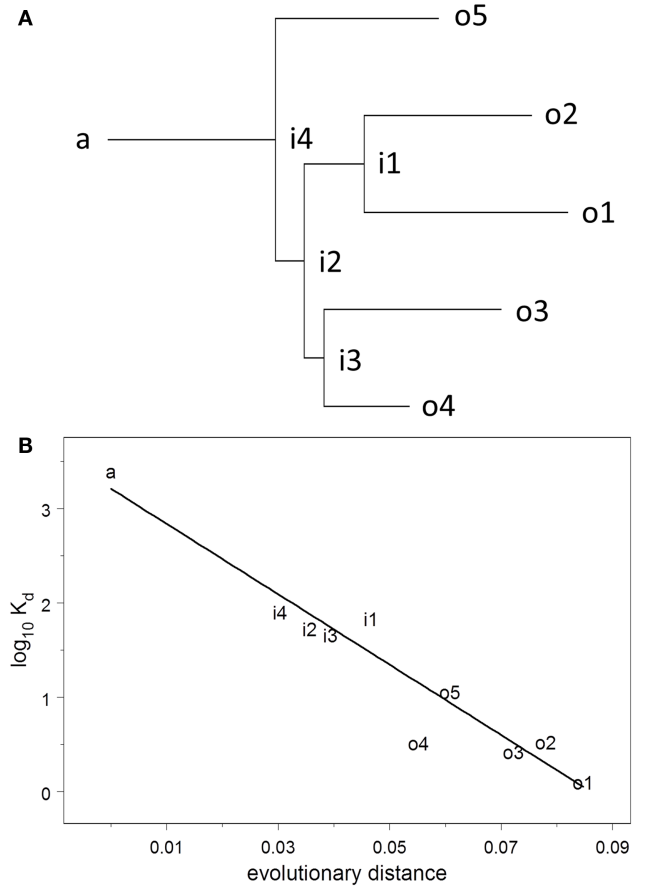


FIGURE 2 | (A) Maximum-likelihood clonal tree showing observed (o), intermediate (i), and ancestral (a) sequences. The tree was inferred using both heavy and light chains. **(B)** Regression analysis of $\log_{10} K_d$ vs. evolutionary distance.

The model we use to estimate parameters and perform tests is straightforwardly derived using likelihood-based methods in statistics. We nevertheless describe the model in some detail below so that the argument may be essentially self-contained.

In order to explore the interplay of selection and mutability, we use a non-linear regression model and multiple independent categorical distributions[1] in which every gene position along each branch of the clonal tree can either be unmutated, mutated synonymously, or mutated non-synonymously. That is, there are three possible classifications for each nucleotide, and the "mutation type" variable takes one of the two values: $T \in \{S, N\}$. For the ith nucleotide in gene g, the variable x_{gi}^T is an indicator for the mutation type. For example, if the nucleotide in question has been mutated non-synonymously along the branch leading up to g from its parent sequence $a(g)$, we have $x_{gi}^N = 1$ and $x_{gi}^S = 0$. If the nucleotide is not mutated at all, we have $x_{gi}^N = 0$ and $x_{gi}^S = 0$.

The relevant likelihood function is the product of independent categorical distributions, whose log (we work with the log of the likelihood function for convenience) is

$$\log L = \sum_{gi} \left[\sum_T x_{gi}^T \log P_{a(g)i}^T + \left(1 - x_{gi}^\bullet\right) \log \left(1 - P_{a(g)i}^\bullet\right) \right] \quad (1)$$

where $P_{a(g)i}^T$ is the probability that the ith nucleotide in the parent of gene g would have mutation type T. The dot in place of an index indicates summation over that index, for example, $1 - P_{a(g)i}^\bullet$ is the probability that the nucleotide in question is not mutated. It is the dependency of these probabilities on the covariates that we model.

The covariates are themselves properties of the specific nucleotide expressed in terms of probabilities. There is first the probability that a given nucleotide mutates at all. This probability is the product of the sequence-specific mutation rate μ_{ai} and the effective evolutionary time τ along the relevant branch. Then, we have the probability σ_{ai}^T that a mutation occurring at the position and gene in question will have type T (that is, conditional on there being a mutation at all). This probability depends on the codon in which the nucleotide is found and its position within the codon. But it also depends on the local sequence (19); these influences have to be estimated for the nucleotide at each position of the gene. Finally, there is the impact of selection. Once a mutation has occurred, it must survive to fixation in order to be observed.

The covariates we will consider for predicting the survival of mutations at the ith position of gene a are the type T of the mutation, the region $R(i) \in$ {FR,CDR} that contains position i, and the mutability μ_{ai}. Note that the dependence of the survival probability on mutability is over and above the role of the mutability in inducing the mutation initially. Indeed, it is the dependence of survival on mutability that is of primary concern for this study. The dependence of survival probability on region is given by the

[1]Such a product distribution is similar to a multinomial distribution, but has different probabilities at each site.

terms γ_R^T. The ratios of these terms give the relative survival probabilities. Because they are introduced as multiplicative rather than additive effects, they are subject, without loss of generality, to the multiplicative constraint $\gamma_{\mathrm{FR}}^{\mathrm{Syn}}\gamma_{\mathrm{CDR}}^{\mathrm{Syn}}\gamma_{\mathrm{FR}}^{\mathrm{Non}}\gamma_{\mathrm{CDR}}^{\mathrm{Non}} = 1$.

Combining all the component probabilities then gives the probability that gene g has acquired an observed mutation of type T at position i and has survived. It is given by

$$P_{\mathrm{gi}}^T = \tau\mu_{a(g)i}\sigma_{a(g)i}^T\left(\gamma_{R(i)}^T + \beta_T\mu_{a(g)i}\right). \quad (2)$$

The local sequence specificity of μ_{ai} and σ_{ai}^T are estimated using external data as described in the supplementary information.

For each hypothesis being tested, we impose the specific constraints on the model parameters in Eq. 2 that correspond to the hypothesis, estimate the remaining parameters by maximizing the likelihood. We then test hypotheses using the likelihood ratio test (20) where applicable, and compare models using the Akaike information criterion (AIC). The AIC is a penalized likelihood, appropriate for model selection where the likelihood ratio test is inapplicable because the respective models are not nested (21).

Local mutability is strongly informative. We compare two models: in the first (Model 0), the mutability is constant over positions $\mu_i = \mu_j$ for all positions i and j. In the second (Model 1), the mutability is determined by the local sequence $\mu_i = m_j$ where m_i is the mutability for the local sequence context at position i, estimated from an independent dataset (see Materials and Methods section). For this test, assume that selection is based on the covariate region × type, and allow γ_R^T to vary subject to the multiplicative constraint above, whereas $\beta_T = 0$ for both T. The models are not nested, so we use AIC and relative likelihood for the comparison. The model with empirical mutability is substantially better supported by the data than is the constant-mutability model (relative likelihood $= 3 \times 10^8$).

Region × type is informative in selection. If region and type are used to classify each potential mutation into one of the four classes that are then used to model the selection process, the predictive power of the model is increased. On comparing the selection-free null model with empirical local mutability (Model 1) with the alternative model in which γ_R^T are fit to the data (Model 2: $\beta_T = 0$, $\mu_i = m_i$), we reject the null model (likelihood ratio test, $p = 0.014$).

Mutability × type is informative in selection. In addition to the mutability that is used to predict the generation of mutations, we may use mutability as a covariate for predicting selection. The resulting model has both linear and quadratic terms in the mutability. On comparing the null model that recognizes type, but not region (Model 3: $\gamma_{\mathrm{FR}}^T = \gamma_{\mathrm{CDR}}^T$, $\beta_T = 0$ and $\mu_i = m_i$), with the alternative model in which β_T are fit to the data (Model 4: $\gamma_{\mathrm{FR}}^T = \gamma_{\mathrm{CDR}}^T$ $\mu_i = m_i$), we reject the null model (likelihood ratio test, $p = 0.010$).

Mutability × type is slightly more informative than region × type in selection. Both region × type and mutability × type have been shown to be predictive. To determine which covariate is more effective as a predictor, we perform a model comparison by AIC; comparing the region × type model (Model 2) with the mutability × type model (Model 4). Both have four degrees of freedom, so by AIC, the comparison favors the mutability × type model (relative likelihood = 1.35).

This result is illustrated in **Figure 3**, which shows the distribution of relative mutabilities in relation to region and the distribution of observed non-synonymous mutations over both gene position and evolutionary time.

The AIC-optimal model uses both mutability × type and region × type to predict mutations. Given the covariates to which we have access, the largest model has $\mu_i = m_i$, and both γ_R^T and β_T are free to vary. This model (Model 5) has the minimum AIC of all models, and all those models that are nested within it are rejected by likelihood ratio tests ($p < 0.05$). The coefficients of the optimal model are shown in **Table 2**.

The selection observed is predominantly purifying. Having determined that selection is measurably occurring, we investigate the nature of the selection by examining the coefficients of the model fit (**Table 2**). In both CDR and FR, the coefficients for non-synonymous mutations are significantly smaller than those for synonymous mutations, consistent with a scenario in which deleterious mutations were introduced in cells that did not survive selection.

Mutability × type is more informative than mutability alone. We have shown that mutability × type is informative. An informative test, the meaning of which will be elaborated on in the discussion, is whether the contribution of mutability to the survival of a mutation depends on the mutation type. For this comparison, we take the null model (Model 6) to have $\beta_{\mathrm{FR}} = \beta_{\mathrm{CDR}}$ and $\gamma_{\mathrm{FR}}^T = \gamma_{\mathrm{CDR}}^T$, and the alternative model (Model 4) with β_T free to vary. The null model is rejected (likelihood ratio test, $p = 8 \times 10^{-3}$).

It is crucial here to understand that this last test is a test of whether type (synonymous vs. non-synonymous) interacts in the statistical sense with mutability (the evolved biases in the targeting of somatic hypermutation) to influence the probability that a mutation survives to fixation. It is taken as given that type alone does influence a mutation's survival probability. It is further taken as given that mutability alone influences whether a mutation occurs in the first place or not. This test is a test of whether mutability is informative regarding the probability that a mutation survives selection. Selection cannot act on synonymous mutations, so evidence that mutability is correlated with selective survival must come from examination of the interaction term between mutability and type. This interaction term is equivalent to $\beta_{\mathrm{FR}} - \beta_{\mathrm{CDR}}$. The rejected null hypothesis is that this quantity is zero.

DISCUSSION

In this study, by inference and expression of the UA and inferred intermediate antibodies of a single clone, we have directly demonstrated the stepwise maturation of antibodies. Such stepwise maturation has been assumed on theoretical grounds (22), but the technology to observe it has not been utilized before now.

The antibodies of clone CL2569 bind influenza HA and are highly mutated. For these reasons, they almost certainly represent a secondary response. In fact, the most likely scenario for the ontogeny of this lineage is that it was formed via affinity maturation during an earlier infection or vaccination

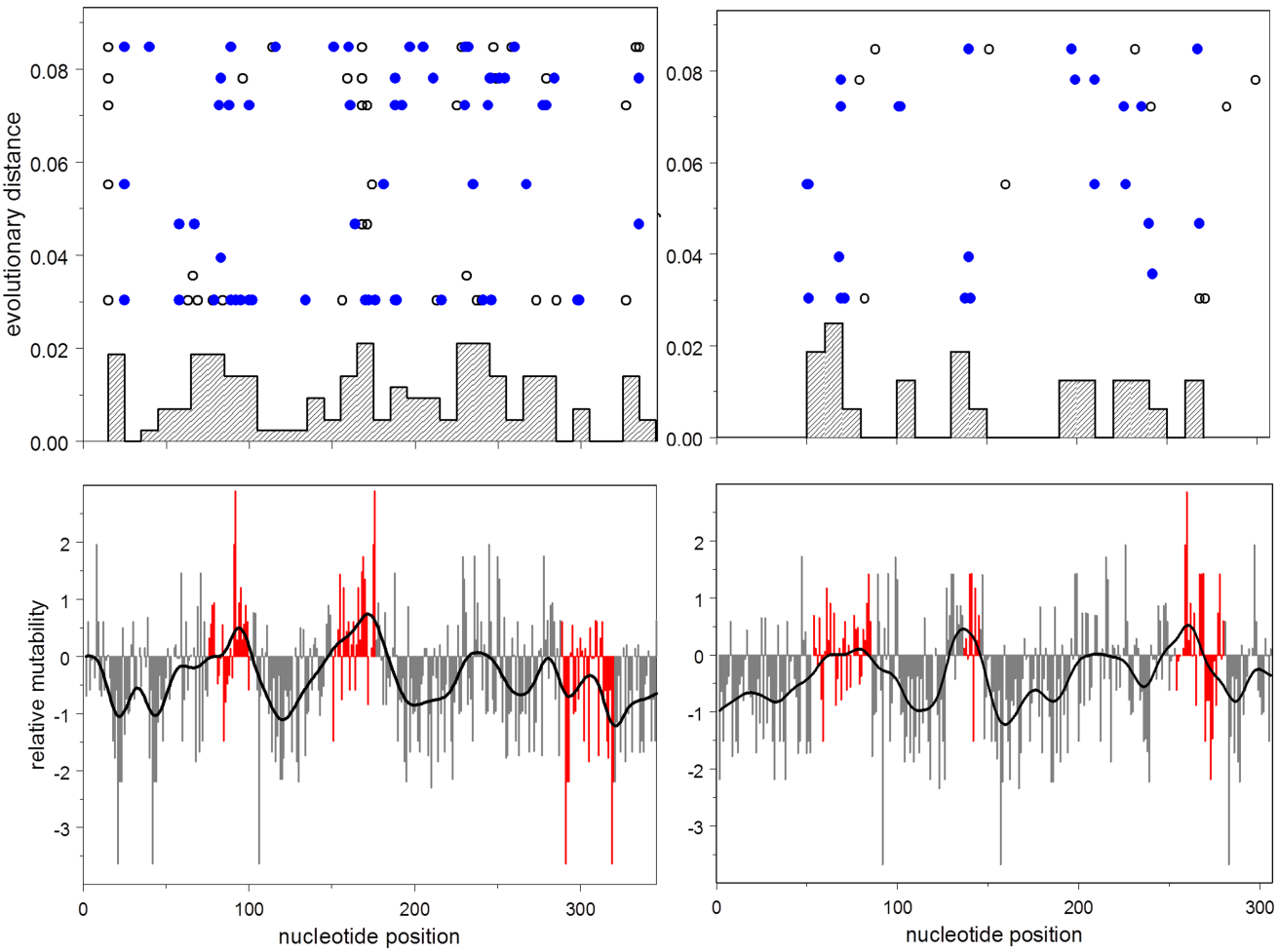


FIGURE 3 | Lower: mutability by position for heavy (left) and light (right) chains. Mutability at CDR nucleotides is shown in red. Upper: histogram by position of accumulated non-synonymous mutations; evolutionary distance vs. position for each synonymous (open disks), and non-synonymous (closed disks) mutation.

Table 2 | Maximum-likelihood estimates for the coefficients in the optimal model.

Model	Mutability	γ_{FR}^{Syn}	γ_{CDR}^{Syn}	γ_{FR}^{Non}	γ_{CDR}^{Non}	β_{Syn}	β_{Non}	AIC
0	Constant	(1)	(1)	(1)	(1)	(0)	(0)	640.0
1	Empirical	(1)	(1)	(1)	(1)	(0)	(0)	589.2
2	Empirical	0.75	2.12	0.67	(0.94)	(0)	(0)	584.6
3	Empirical	1.18	(1.18)	(0.85)	(0.85)	(0)	(0)	589.2
4	Empirical	2.25	(2.25)	(0.44)	(0.44)	−25.8	16.6	584.0
5	Empirical	1.58	3.23	0.34	(0.58)	−21.4	13.8	579.0
6	Empirical	1.29	(1.29)	(0.78)	(0.78)	9.63	(9.63)	589.0

Parentheses indicate that the parameter is invariant at the indicated value in the model considered.

and was subsequently activated into differentiation to plasmacytes by the experimental infection without undergoing further affinity maturation. The subject was infected with the H3N2 A/Wisconsin/67/2005 strain of influenza virus; preliminary binding assays were done on HA from several strains including the infecting strain, H1 A/Brisbane/59/2007, and several others. Although the maturation patterns were similar across several of the strains, the affinities measured against H1 A/Brisbane/59/2007 were generally higher (15). The infection study was performed in 2008, so previous infection in the subject with influenza strains circulating in 2007 is consistent with this observed reactivity to H1 A/Brisbane/59/2007.

The recovered mAb in this clonal lineage were mostly IgM with a single member that was IgA1, and all the members had a degree of somatic hypermutation consistent with one or more prior rounds of antigen-driven germinal center maturation. Recent work by Pape et al. (23) has shown that in mice IgM-memory B-cells and class-switched memory B-cells have different circulation kinetics, such that IgM-memory B-cells persist after class-switched memory B-cells have disappeared from circulation. Furthermore, upon restimulation with antigen, IgM-memory B-cells were less likely to produce a secondary response in the presence of antigen-specific plasma antibody. Thus, it is interesting that the members

of this clonal lineage bind to various previously circulating strains including the older H3 A/Johannesburg/33/1994 strain, that the antibodies were predominantly IgM, were hypermutated, and did not significantly contribute to the plasma antibody pool 4 weeks after experimental infection (15). All these findings suggest that this lineage is an example of such an IgM-memory B-cell clone isolated from an influenza-infected human subject.

Like other Darwinian processes, affinity maturation arises in the interplay between the generation of diversity and the subsequent selection of fitter variants. Affinity maturation, however, is a somatic process; properties of the germline gene segments that facilitate efficient maturation are preserved for the next germline generation (8). Thus, mutation and selection in affinity maturation are very strongly intertwined with mutations that are more likely on average to confer advantage, produced more frequently than those that are more likely on average to confer disadvantage. This circumstance has a practical consequence, complicating the analysis of selective pressure. We have overcome that problem by estimating the relevant characteristics of somatic hypermutation from a collection of human heavy chain genes rearranged out of frame and insusceptible to selection.

SELECTION AND MUTABILITY SYNERGIZE DURING AFFINITY MATURATION TO HA

The local codon bias that is present in Ig V-gene segments and increases mutability in the CDR creates a strongly non-uniform probability distribution over the links between Ig genes in the genotype space (**Figure 4**). Each of the Ig genes at the nodes of this space has an effective affinity for the antigen HA associated with it, which presumably determines the relative fitness of B-cells expressing the antibody encoded by that gene. Because of the mutational bias, from any starting node there are preferred nodes, which are visited with greater probability and in less time on average, than others. The question addressed here is whether the sequences more likely to be visited during somatic hypermutation because of this bias are also more likely to encode antibodies that confer a selective advantage.

Figure 4 is a simple cartoon intended to illustrate the idea. The grid represents the genotype space (although the topology is not at all realistic). The dark arrows indicate the directions of preferred mutations. We consider the node 1 to be the starting node. The other nodes 2–4 are each six mutations away from node 1, but they differ in the number of non-preferred mutations that are required to reach them. In the real system, we can estimate the mutation rate for each link, and in particular can estimate the mutation rates over the links connecting the nodes actually occupied during affinity maturation. We also have measured the affinity at each of these nodes, and know that they represent increases over time. So the question is, "are the visited nodes largely close to the preferred paths (as are nodes 2 and 3 in **Figure 4**), or randomly placed with respect to the preferred paths (illustrated by node 4)?"

We expect that such correlation between mutational preference and selective advantage holds on average over the history of antigens encountered by the gene segment in question. It is hypothesized that this is the reason why such local codon bias exists in the first place. The question addressed here is whether such a correlation exists, not on average, but in this particular instance, for this one specific antigen.

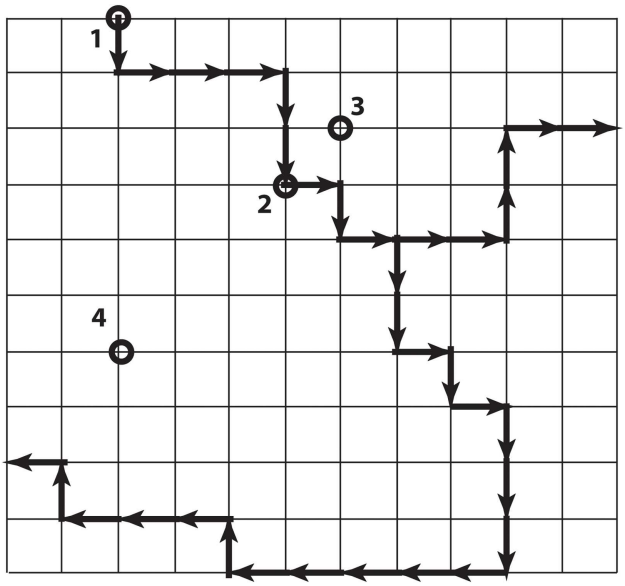


FIGURE 4 | Simplified illustration of genotype space with preferred directions. Each node is a DNA sequence, and neighbors differ by one nucleotide. The dark arrows show preferred directions, meaning the mutation along the direction of the arrow occurs at a higher rate than mutations along the regular paths. The nodes labeled 2, 3, and 4 are all six steps from node 1, but differ in the number of non-preferred steps that must be taken to arrive there from 1.

The mutability is defined at each nucleotide position as the probability of a mutation at that position conditional on there being exactly one mutation in the gene, and no selection on the gene product. In the presence of selection, the probability that a mutation will be fixed is the product of the probability that the mutation occurs at all and the probability that, once it has occurred, it is preserved through selection. The hypothesis we are testing is that the second of these probabilities, the probability of preservation, is itself functionally depending on the mutability. The order of the causality would be that the mutabilities have been adjusted, largely through codon usage, to make evolution toward the potentially advantageous genes more rapidly and more reliably.

We address the question in **Figure 5**, which shows the empirical cumulative distribution plots of synonymous and non-synonymous mutations as a function of mutability, compared to three theoretical models: zero order (mutability has no influence, even on the probability of having a mutation in the first place), first order (mutability has the influence expected under selection-free conditions), and second-order (probability of selection is directly proportional to mutability). The plots show that the synonymous mutations are consistent, as expected, with the first-order model. Indeed, this plot should be regarded as a test of the accuracy of the estimated mutability, which appears to be adequate, although the mutabilities of the higher-mutability positions may be somewhat over-estimated. In contrast, the observed non-synonymous mutations fall between the first- and second-order curves, consistent with synergy between local codon bias and selection. **Figure 5** is merely suggestive; the direct test of the relevant hypothesis (Model 4 vs. Model 6) provides stronger evidence.

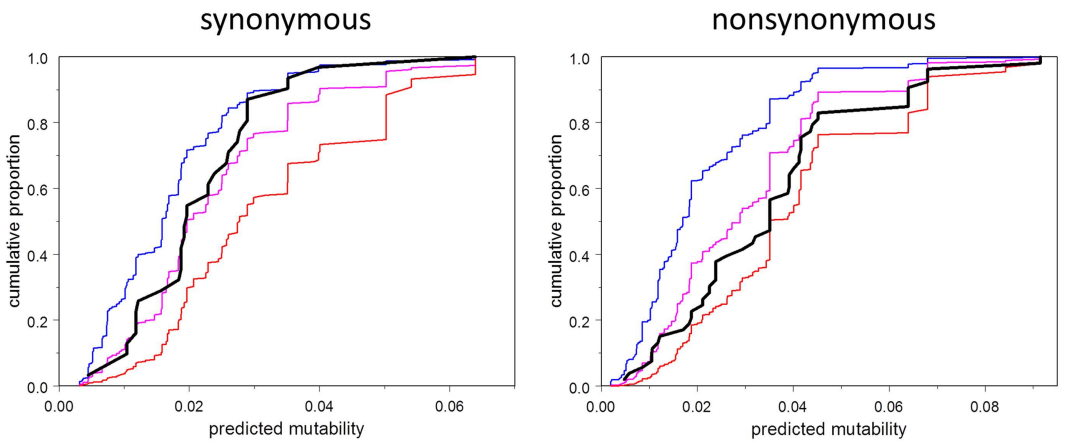


FIGURE 5 | Cumulative distribution function (CDF) of mutability among observed mutations (black), and corresponding to three models: order 0 (no effect of mutability at all, blue), order 1 (consistent with selection random with respect to mutability, magenta), second order (selection proportional to mutability, red). Note that the observed CDF for synonymous mutations is approximately consistent with the order one model, and falls between the order zero and order one curves in any case. The CDF for non-synonymous mutations falls between the order one and order two curves.

This test says that the influence of mutability on the survival of a mutation depends on the type of mutation, whether synonymous or non-synonymous. If the mutation is non-synonymous, the mutability has greater positive predictive power than that of synonymous type.

CONCLUSION

Strikingly, despite the fact that the dissociation constant changed by three orders of magnitude from the common ancestor to the observed mature antibodies, the distribution of mutations is heavily biased toward those with high intrinsic mutability, suggesting that selection worked in synergy with local codon bias in the maturation of CL2569. This analysis suggests that affinity maturation is strongly constrained to occur by mutational diffusion along preferred paths in genotype space, with selection acting negatively on genotypes in this network that fail to confer enhanced antigen-binding affinity. There is no evidence for selection pulling the evolving clone substantially out of the mutationally preferred paths.

There are many highly effective vaccines that work through the induction of a potent humoral response, but there are many devastating infectious diseases for which no effective vaccine is yet available in spite of intense research efforts, including malaria, hepatitis C, and HIV-1. The agents of these diseases do not typically elicit protective natural immunity, so new approaches to vaccine development may be indicated. One such approach is predicated on the observation that the efficiency of immunogen stimulation of germinal center naïve and intermediate B-cell antibodies is determined by immunogen affinity for B-cell precursor B-cell receptor (24–26). Design of immunogens with high-affinity binding for antibody UAs and their intermediates is now possible with the computational methods described in this study (27). It is our hope that the emerging understanding of the intertwined mechanisms of diversification and selection in affinity maturation will open new avenues for vaccine engineering.

MATERIALS AND METHODS

STATISTICAL AND COMPUTATIONAL

All analyses and computational manipulations were performed using software developed in the Kepler laboratory.

ELISA data analysis

The data from the ELISA dilution series were fit to a Hill function with Hill coefficient = 1 and additive background (28). The maximum value of the optical density and the value of the background optical density were taken to be equal over all wells on a given plate.

Inference of unobserved antibodies: ancestral rearrangement and maturation intermediates

We compute the posterior probability mass function on the nucleotides at each position of the unmutated common ancestor given the set of clonally related observed Ig genes of CL2569, as described in detail in Ref. (13).

Inference of somatic hypermutation sequence specificity

We searched NCBI Genbank for rearranged human Ig heavy-chain variable-region genes and retrieved and validated 34,546 genes. We eliminated genes with possible clonal relatives in the set by randomly eliminating all but one of each sequence within groups likely to be clonally related. Two antibodies were considered likely to be related if they shared the same inferred IGHV and IGHJ genes (without regard to allele) and shared at least 75% nucleotide identity in CDR3[2]. From these, we selected those that were likely to have been rearranged out of frame as evidenced by the number of

[2]This is admittedly a crude estimation procedure, but sequence set is small enough that we expect few if any errors from its use. Furthermore, we are unconcerned about falsely excluding unrelated sequences, which is the only likely error to be made by this method. Finally, the proper statistical procedure for testing clonal relatedness is sufficiently complex (Thomas B. Kepler, in preparation) that to put aside the space in a paper that does not require its full power would be distracting.

nucleotides between the intact invariant cysteine in VH FR 3 and the intact invariant tryptophan in JH being other than a multiple of three.

By counting frame-shift mutations in the VH-encoded part of the gene, which have resulted from somatic mutations or sequencing error, we estimate the likely number of genes that would have frame-shift mutations in CDR3 to be about 195 genes. That is about 11% of our candidate non-productively rearranged genes are likely to have been rearranged in-frame and to have acquired their frame-shift mutations subsequently.

To ameliorate the impact this contamination could have on the downstream analysis, we removed all genes inferred to have been rearranged to a VH1 family member. The reason for this filtering step is that the positions of pentanucleotides in the remaining sequences will be significantly de-correlated from the positions of the corresponding pentanucleotides in the target sequences, which are rearranged to a VH1 family member.

After this filtering step, 1707 sequences remained, containing 9961 nucleotide substitutions in 423,654 total bases.

The mutation frequency for the central position at each pentanucleotide motif was computed by scanning each inferred UA. Of the 4^5 (1024) possible pentanucleotides, 938 motifs were present in the total dataset, 922 in the out of frame dataset. Of the motifs with at least 100 observations among the UAs in the non-productive set, 24 of them had no mutations. In contrast, the motif AGCTA, which is consistent with the canonical "hot-spot" RGYW, was mutated at the center position 112 out of 618 times for a frequency of 18%.

For comparison to other such datasets previously assembled, we also computed the trinucleotide mutation frequencies. The spearman correlation between our trinucleotide mutation frequencies and the corresponding mutability indices from unselected sequences in the study by Shapiro et al. (29) is 0.80, indicating a high level of agreement between the two sets.

Rather than use, the raw count ratios for the mutability and mutation spectrum estimates directly (which is likely to result in over- or under-fitting), we chose to fit these data to a variable-motif length model using regression trees. The first statistical treatment of sequence specificity in somatic mutation produced hot-spot motifs of different lengths (7) and it seems natural to fit such a model now that much more data are available.

The end result of this estimation procedure is a set of nucleotide motifs that are mutually exclusive and complete (every nucleotide in any DNA sequence will belong to exactly one motif) to each member of which is assigned a mutation rate. Each motif may be up to 5 nt long. The procedure is as follows.

Each node in the decision tree contains a pentanucleotide motif of the form n_1, n_2, n_3, n_4, n_5 in which each $n_i = \{$A, G, T, C, R, Y, S, W, N$\}$ where R, Y, S, W, N are the IUPAC symbols respectively for purine (A or G), pyrimidine (T or C), weak (A or T), strong (G or C), and any (A, G, T, or C) nucleotide.

The function to be maximized, the objective function, is the log of the marginal likelihood summed over all nodes in the tree. The overall likelihood is the product of the binomial likelihoods at each node. At each node, the prior distribution on mutations is a beta distribution with parameters $\alpha = 1, \beta = 47$. The beta distribution is chosen because it is conjugate to the binomial distribution, and the specific parameters are chosen because they maximize the information entropy at the observed average mutation frequency in the set, 2.1%. As such, this prior is the most uninformative prior consistent with the average mutation frequency.

The marginal likelihood for a node with m mutations and u unmutated bases is computed by integrating over the mutation probabilities in the product of the likelihood and prior density functions giving:

$$L_{\mathrm{MU}}(m, u|\alpha, \beta) = \frac{\Gamma(\alpha + m)\,\Gamma(\beta + u)\,\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha)\,\Gamma(\beta)\,\Gamma(\alpha + \beta + m + u + 1)} \quad (3)$$

where Γ is the gamma function.

The tree-building algorithm is greedy, choosing the best available split at each node. Allowed splits at any step in the algorithm at any single position in the motif are as follows:

$$N \to R/Y, N \to S/W, R \to A/G, Y \to T/C,$$
$$S \to G/C, W \to A/T.$$

This scheme ensures that each pentanucleotide is mapped to exactly one terminal node on the tree at all stages of the procedure. A node is declared terminal if the product of the marginal likelihoods for the two daughter nodes in the optimal split is less than the marginal likelihood of the parent node, that is, if the likelihood cannot be increased any further at that node.

The result of applying this process to our count data is a tree with 55 terminal nodes. Among these, the one with the lowest relative mutability is with YTGGS with posterior mean $\widehat{p} = 7.9 \times 10^{-4}$. The AID "hot-spot" motif AGCT is assigned to the NAGCW node, with $\widehat{p} = 8.8 \times 10^{-2}$.

Regression model for the dependence of selection on mutability

The model scheme for analysis of selection is described in the main text. The data used are all nucleotide position in the heavy-chain variable regions up to and including the nucleotides of the FR3 invariant cysteine codon. The fitting of parameters by maximum likelihood was performed by numerical optimization using the Nelder–Mead simplex algorithm using a software implementation based largely on that described in Numerical Recipes (30).

Statistical hypothesis tests are based on the likelihood ratio test when models are nested. Model comparison is done by differential AIC expressed as relative likelihoods (31).

EXPERIMENTAL

Clinical protocol

The clinical EI protocol study was performed at Retroscreen Virology Ltd. (Brentwood, UK) as previously described (32) using a protocol approved by their local ethics board and the Duke IRB. Subjects were prescreened and provided informed consent before being given a nasal challenge with influenza A/Wisconsin/67/2005 (H3N2) challenge stock manufactured under current good manufacturing practices by Baxter BioScience (Vienna, Austria). Intranasal challenge was given using $10^{3.08}$ TCID50 to subject EI13 from whom the antibodies described in this study were derived. In this protocol, blood was drawn before challenge, then daily on days 0–7, and on day 28 after challenge. Symptoms were recorded twice daily using a modified Jackson scoring system (33). Productive infection was confirmed by active viral shedding detected by assays of nasal washes obtained during the 7-day quarantine period.

Single-cell flow cytometry sorting strategy

Human peripheral blood mononuclear cell samples collected 7 days after infection with A/Wisconsin/67/2005 (H3N2) were labeled with panels of fluorochrome-antibody conjugates specific for human CD3 (PE-Cy5), CD16 (PE-Cy5), CD19 (APC-Cy7), CD20 (PE-Cy7), CD27 (Pacific Blue), CD235a (PE-Cy5), IgD (PE), IgM (FITC) (all, BD Biosciences, San Jose, CA, USA), CD14 (PE-Cy5), and CD38 (APC-Cy5.5) (both Invitrogen, Carlsbad, CA, USA). Plasma cells/plasmablasts were sorted into 20 μl/well RT/PCR buffer in 96-well plates as described (10, 12) by gating on $CD3^{-}$ $CD14^{-}$ $CD16^{-}$ $CD235a^{-}$ $CD19^{+}$ $CD20^{-/lo}$ $CD27^{hi}$ $CD38^{hi}$ cells. All antibody reagents were tittered and used at optimal concentrations for flow cytometry.

PCR amplification of plasmablast/plasma cell immunoglobulin VH and VL variable-region genes

The Ig VH and VL variable-region genes of the sorted plasmablast were amplified by RT and nested PCR using the method as reported (11). The PCR products amplified by this method contain enough coding region sequences for the constant regions of either heavy- or light-chain genes for allowing the identification of IgH subclass and light-chain types (12). Isolated VH and VL variable-region genes were used to assemble full-length Ig IgG1 heavy- and light-chain expression cassette by overlapping to express recombinant IgG1 antibodies using the method as described (12).

Expression of VH and VL variable-region genes as IgG1 recombinant mAb

The isolated Ig VH and VL gene pairs were assembled by PCR into the linear full-length Ig heavy- and light-chain gene expression cassettes for production of recombinant mAbs by transfection in the human embryonic kidney cell line, 293T (ATCC, Manassas, VA, USA) using the methods as described (12). The purified PCR products of the paired Ig heavy- and light-chain gene expression cassettes were co-transfected into near confluent 293T cells grown in 6-well (2 μg of DNA for each cassettes per well) tissue culture plates (Becton Dickson, Franklin Lakes, NJ, USA) using PolyFect (Qiagen, Valencia, CA, USA) or Effectene (Qiagen Valencia, CA, USA) using protocols recommended by the manufacturers. Six to eight hours after transfection, the 293T cells were fed with fresh culture medium supplemented with 2% FCS and were incubated at 37°C in a 5% CO_2 incubator. Culture supernatants were harvested 3 days after transfection and quantified for expressed IgG levels and screened for antibody specificity.

Antibodies that bound HA in a screening assay as well as the inferred UA and intermediate clonal antibodies were produced on a larger scale so that screening assays could be replicated and broadened to more fully define the range of binding activity of expressed plasma cell derived-antibodies. Purified recombinant antibodies were produced in bulk cultures by transient transfection using Ig heavy- and light-chain genes cloned in pcDNA plasmids (12). The Ig heavy- and light-chain gene expression cassettes used for production of recombinant antibodies for initial screening were cloned into pcDNA 3.3 (Invitrogen, Carlsbad, CA, USA) for production of purified recombinant mAbs using standard molecular protocol, and co-transfected into 293T cells cultured in T175 flasks using PolyFect (Qiagen, Valencia, CA, USA) or polyethylenimine (34), cultured in DMEM supplemented with 2% FCS. Recombinant mAbs were purified from culture supernatants of the transfected-293T cells using anti-human Ig heavy-chain-specific antibody–agarose beads (Sigma, St. Louis, MO, USA) using the method as previously described (12, 34). Purified antibodies used in the study were confirmed having typical patterns of predominant whole IgG in SDS-PAGE and Western blots under reducing and non-reducing conditions (12).

Binding antibody multiplex assay

Concentration of recombinant mAbs secreted in the transfected-293T cell culture in the supernatants was determined using a method previously described (12). The expressed recombinant mAb were assayed for antibody reactivity by a standardized binding antibody multiplex assay (35) performed in a GCLP compliant manner. Binding specificities to influenza vaccine 2007 (Fluzone® 2007), trivalent influenza vaccine 2008 (Fluzone® 2008), and baculovirus-derived HA proteins (H1N1 A/Brisbane/59/2007, H1N1 A/California/04/2009, H1N1 A/Solomon Islands/03/06, H3N2 A/Brisbane/10/2007, H3N2 A/Johannesburg/33/1994, H3N2 A/Johannesburg/33/1994, H3N2 A/Wisconsin/67/05, B/Florida/04/06; Protein Sciences, Meriden, CT, USA) were determined using purified mAb diluted serially starting at 50 μg/ml.

ELISA data analysis for estimation of K_d

Purified mAb prepared at known concentrations were evaluated by ELISA against baculovirus-expressed purified hemagglutinin (H1 A/Brisbane/59/2007; Protein Sciences, Meriden, CT, USA). Samples were diluted serially for the analysis and data were analyzed using the model

$$y_i = \log\left[\alpha + (\beta - \alpha)\frac{c_i}{K_{\mathrm{d}} + c_i}\right] + \varepsilon_i \qquad (4)$$

where y_i is the log of the optical density measured at the ith dilution, α is the background optical density, β is the maximum optical density, K_{d} is the equilibrium dissociation constant, c_i is the known concentration of analyte at the ith dilution, and the ε are independent, identically distributed Gaussian errors. For each antibody studied, the parameters of this model were fit using software developed for the purpose (28).

ACKNOWLEDGMENTS

We are grateful for fruitful discussions with the members of the Center for Computational Immunology, and the members of the Duke Human Vaccine Institute's Antibodyome research group. Grant Information: this work was supported by NIH/NIAID research contract HHSN272201000053C (to Thomas B. Kepler, PI) and a Vaccine Development Center grant in the Collaboration for AIDS Vaccine Discovery Program from the Bill and Melinda Gates Foundation (Barton F. Haynes, PI).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00170/abstract

Datasheet 1 | Sequence alignment for CL2569 heavy chain, including observed and inferred sequences.

Datasheet 2 | Sequence alignment for CL2569 light chain, including observed and inferred sequences.

Datasheet 3 | Tables of results for sequence specificity of mutation frequency.

REFERENCES

1. Yoshida T, Mei H, Dörner T, Hiepe F, Radbruch A, Fillatreau S, et al. Memory B and memory plasma cells. *Immunol Rev* (2010) **237**:117–39. doi:10.1111/j.1600-065X.2010.00938.x
2. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* (2007) **76**:1–22. doi:10.1146/annurev.biochem.76.061705.090740
3. Brown LE, Murray JM, White DO, Jackson DC. An analysis of the properties of monoclonal antibodies directed to epitopes on influenza virus hemagglutinin. *Arch Virol* (1990) **114**:1–26. doi:10.1007/BF01311008
4. Usinger WR, Lucas AH. Avidity as a determinant of the protective efficacy of human antibodies to pneumococcal capsular polysaccharides. *Infect Immun* (1999) **67**:2366–70.
5. Lambert P-H, Liu M, Siegrist C-A. Can successful vaccines teach us how to induce efficient protective immune responses? *Nat Med* (2005) **11**(4 Suppl):S54–62. doi:10.1038/nm1216
6. Delgado MF, Coviello S, Monsalvo AC, Melendi GA, Hernandez JZ, Batalle JP, et al. Lack of antibody affinity maturation due to poor toll-like receptor stimulation leads to enhanced respiratory syncytial virus disease. *Nat Med* (2009) **15**:34–41. doi:10.1038/nm.1894
7. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta* (1992) **1171**:11–8. doi:10.1016/0167-4781(92)90134-L
8. Kepler TB. Codon bias and plasticity in immunoglobulins. *Mol Biol Evol* (1997) **14**:637–43. doi:10.1093/oxfordjournals.molbev.a025803
9. Wagner SD, Milstein C, Neuberger MS. Codon bias targets mutation. *Nature* (1995) **376**:732–732. doi:10.1038/376732a0
10. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, Larsen C, et al. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* (2008) **453**:667–71. doi:10.1038/nature06890
11. Tiller T, Meffre E, Yurasov S, Tsuiji M, Nussenzweig MC, Wardemann H. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *J Immunol Methods* (2008) **329**:112–24. doi:10.1016/j.jim.2007.09.017
12. Liao HX, Levesque MC, Nagel A, Dixon A, Zhang R, Walter E, et al. High-throughput isolation of immunoglobulin genes from single human B cells and expression as monoclonal antibodies. *J Virol Methods* (2009) **158**:171–9. doi:10.1016/j.jviromet.2009.02.014
13. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res* (2013) **2**:103. doi:10.12688/f1000research.2-103.v1
14. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* (2003) **301**:1374–7. doi:10.1126/science.1086907
15. Moody MA, Zhang R, Walter EB, Woods CW, Ginsburg GS, McClain MT, et al. H3N2 influenza infection elicits more cross-reactive and less clonally expanded anti-hemagglutinin antibodies than influenza vaccination. *PLoS One* (2011) **6**(10):e25797. doi:10.1371/journal.pone.0025797
16. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res* (2009) **37**:D1006–12. doi:10.1093/nar/gkn838
17. Kreitman M, Akashi H. Molecular evidence for natural selection. *Annu Rev Ecol Syst* (1995) **26**:403–22. doi:10.1146/annurev.es.26.110195.002155
18. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) **20**:683–94. doi:10.1093/intimm/dxn026
19. Cowell LG, Kepler TB. The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation. *J Immunol* (2000) **164**:1971–6.
20. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A* (1933) **231**:289–337. doi:10.1098/rsta.1933.0009
21. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* (1974) **19**:716–23. doi:10.1109/TAC.1974.1100705
22. Kepler TB, Perelson AS. Somatic hypermutation in B cells: an optimal control treatment. *J Theor Biol* (1993) **164**:37–64. doi:10.1006/jtbi.1993.1139
23. Pape KA, Taylor JJ, Maul RW, Gearhart PJ, Jenkins MK. Different B cell populations mediate early and late memory during an endogenous immune response. *Science* (2011) **331**:1203–7. doi:10.1126/science.1201730
24. Dal Porto JM, Haberman AM, Kelsoe G, Shlomchik MJ. Very low affinity B cells form germinal centers, become memory B cells, and participate in secondary immune responses when higher affinity competition is reduced. *J Exp Med* (2002) **195**:1215–21. doi:10.1084/jem.20011550
25. Shih TA, Meffre E, Roederer M, Nussenzweig MC. Role of BCR affinity in T cell dependent antibody responses in vivo. *Nat Immunol* (2002) **3**:570–5. doi:10.1038/ni776
26. Schwickert TA, Victora GD, Fooksman DR, Kamphorst AO, Mugnier MR, Gitlin AD, et al. A dynamic T cell-limited checkpoint regulates affinity-dependent B cell entry into the germinal center. *J Exp Med* (2011) **208**(6):1243–52. doi:10.1084/jem.20102477
27. Haynes BF, Kelsoe G, Harrison SC, Kepler TB. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat Biotechnol* (2012) **30**:423–33. doi:10.1038/nbt.2197
28. Feng F, Sales AP, Kepler TB. A Bayesian approach for estimating calibration curves and unknown concentrations in immunoassays. *Bioinformatics* (2011) **27**:707–12. doi:10.1093/bioinformatics/btq686
29. Shapiro GS, Aviszus K, Murphy J, Wysocki LJ. Evolution of Ig DNA sequence to target specific base positions within codons for somatic hypermutation. *J Immunol* (2002) **168**:2302–6.
30. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge University Press (2007).
31. Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer (2002). p. 49–97.
32. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO3rd, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe* (2009) **6**:207–17. doi:10.1016/j.chom.2009.07.006
33. Jackson GG, Dowling HF, Spiesman IG, Boand AV. Transmission of the common cold to volunteers under controlled conditions. I. The common cold as a clinical entity. *AMA Arch Intern Med* (1958) **101**:267–78. doi:10.1001/archinte.1958.00260140099015
34. Smith K, Garman L, Wrammert J, Zheng NY, Capra JD, Ahmed R, et al. Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat Protoc* (2009) **4**:372–84. doi:10.1038/nprot.2009.3
35. Tomaras GD, Yates NL, Liu P, Qin L, Fouda GG, Chavez LL, et al. Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *J Virol* (2008) **82**:12449–63. doi:10.1128/JVI.01708-08

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2013; accepted: 30 March 2014; published online: 22 April 2014.

*Citation: Kepler TB, Munshaw S, Wiehe K, Zhang R, Yu J-S, Woods CW, Denny TN, Tomaras GD, Alam SM, Moody MA, Kelsoe G, Liao H-X and Haynes BF (2014) Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. Front. Immunol. **5**:170. doi: 10.3389/fimmu.2014.00170*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Kepler, Munshaw, Wiehe, Zhang, Yu, Woods, Denny, Tomaras, Alam, Moody, Kelsoe, Liao and Haynes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

OPINION ARTICLE
published: 26 May 2014
doi: 10.3389/fimmu.2014.00237

A major hindrance in antibody affinity maturation investigation: we never succeeded in falsifying the hypothesis of single-step selection

Michal Or-Guil [1,2] **** and Jose Faro*** [3,4,5] *******

[1] *Systems Immunology Laboratory, Department of Biology, Humboldt-Universität zu Berlin, Berlin, Germany*
[2] *Research Center ImmunoSciences, Charité-Universitätsmedizin Berlin, Berlin, Germany*
[3] *Area of Immunology, Faculty of Biology, Biomedical Research Center (CINBIO), Universidade de Vigo, Vigo, Spain*
[4] *Instituto Biomédico de Vigo, Vigo, Spain*
[5] *Instituto Gulbenkian de Ciência, Oeiras, Portugal*
**Correspondence: michal.orguil@gmail.com; jfaro@uvigo.es*

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Joshy Jacob, Emory University, USA

Keywords: antibody affinity maturation, somatic hypermutation, *V* gene sequences, phylogenetic trees, selection mechanism

INTRODUCTION

Antibody (Ab) affinity maturation (AAM) referred originally to the observed increase in average Ab affinity against a hapten (1). Later, it was found that AAM is associated with the formation of transient lymphoid structures in the B cell zones of lymphoid tissues, called germinal centers (GC), during T-cell dependent immune responses in higher vertebrates (2).

In another line of research, AAM was related to the occurrence of mutations in the variable (V) domain of Ab heavy (H) and light (L) chains, respectively, V_H and V_L. In those works, a mutational analysis of Ab V genes was performed, initially on bulk splenic plasma B cells and later on GC B cells vs. extrafollicular B cells, after successive immunizations. The results showed typically an increased number of mutated GC B cells (3–6), and an accumulation of mutations per Ab chain during the ongoing immune response, with many mutated B cells displaying higher affinity for the hapten used for immunization. This provided strong support to a previously suggested concept (7), according to which AAM is a B-cell receptor (BCR)-based Darwinian evolutionary process.

A few years later, two complementary hypotheses were proposed. The first one, based on a mathematical model, suggested that, for the fastest production of high affinity Abs, the mutation rate in GC B cells should be minimal before GCs reach a threshold size, and then switch abruptly to the maximal possible rate (8). The second hypothesis proposed, for the assumed Darwinian process, alternating cycles of B cell proliferation plus mutation plus selection (9). These ideas were soon extended in another modeling work, showing that Ab affinity can be maximized when the mutational mechanism switches on and off regularly (10). These results contributed considerably to strengthen the general belief in the recycling or multiple-step selection hypothesis. On the other hand, more recently, alternative B cell selection mechanisms were proposed that do not require multiple-step selection in order to be compatible with observed levels of Ab affinity increase during a primary immune response (11, 12).

There is still much to learn about AAM mechanisms (13–17), and there is a need to clarify some aspects of the GC physiology where overinterpretation and preconceptions prevail (18, 19). The multiple-step selection hypothesis is a prominent example of a concept that, having important basic and practical implications, has never been confirmed. Clearly, a direct way to establish it would be to observe multiple BCR-mediated selection events by tracking individual B cells via imaging of lymphatic tissue, observing SHM taking place between selection rounds. However, direct observation of even one selection event is not yet possible. At the same time, attempts to interpret indirect data must be faulty due to the need to use unverified assumptions on AAM mechanisms.

Therefore, we take here a radically different approach: we propose to consider the single-step selection concept to be a null-hypothesis which should be attempted to be falsified (**Figure 1**). Because this ansatz puts the focus on a process of random non-directed acquisition of mutations, it minimizes the need for unverified assumptions. And because mutations carry the signature of the selection process, the data to be used should consist of Ab V gene sequences. In the following, we examine two possible falsifying strategies.

FALSIFYING THE NULL-HYPOTHESIS WITH PHYLOGENETIC TREES

Let us consider all mutated V_H or V_L sequences belonging to a given B cell lineage. The corresponding phylogenetic tree is a result of the evolutionary process undergone by the initial sequence, and as such, is shaped by the various factors pertaining to the affinity maturation process.

Extensive work was performed on developing methods to build phylogenetic trees from V genes of a common lineage (20) and to analyze how shape measures depend on AAM mechanisms (21, 22). These simulations show that the tree shapes vary most on the initial clone affinity and the selection threshold, and much less in dependence on the rates of GC B cell recycling (22), not allowing for a unique mapping from tree shapes to selection mechanisms – likely

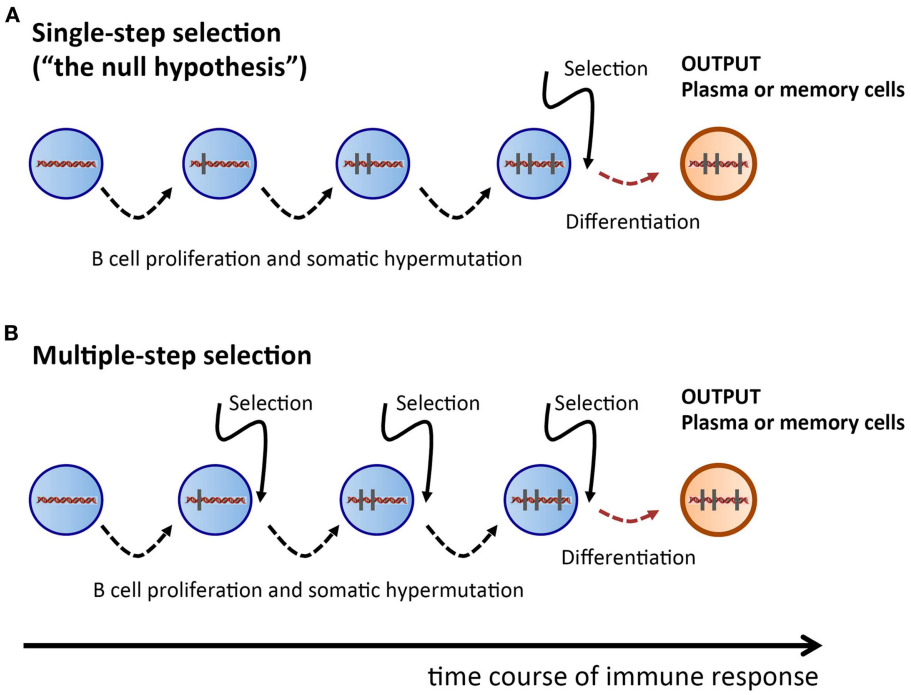


FIGURE 1 | Sketch of proliferation plus SHM and selection history of a plasma or memory cell. **(A)** Single-step selection. Several cell division plus mutation cycles and a single final selection step before terminal differentiation into a plasma or memory cell. **(B)** Multiple-step selection. Several alternating rounds of cell division plus mutation and selection [corresponding to several rounds of **(A)**], followed by terminal differentiation into a plasma or memory cell after the last selection step. Vertical bars indicate mutations.

because the investigated trees were small. In addition to global measures not always being helpful in pointing to mechanisms at the micro-evolutionary scale (15, 23, 24), simulation of global measures like peak total GC B cell numbers did not lead to results that contradict the single-step hypothesis (22).

Summing up, the null-hypothesis has never been falsified by examining the shapes of phylogenetic trees.

FALSIFYING THE NULL-HYPOTHESIS BY COUNTING RECURRENT MUTATIONS

Let us consider a thought experiment, in which two syngeneic mice with a single, non-mutated B cell clone expressing the same V genes, are immunized with the same antigen, and that both mice initiate a process of AAM in which the selection forces acting on the diversified V sequences are identical. Let us further assume that the baseline mutability during SHM is uniformly distributed along rearranged V genes and independent on the time elapsed after immunization. After a number of days, a sample of Ab V genes is sequenced and an *independent set of V sequences* is obtained for each mouse. As a result of the stochastic nature of SHM, the mutation distribution in both sets may be quite different. If nevertheless an identical set of mutations appears in both independent data sets, we call it a *recurrent mutation pattern*.

How likely is it to find a recurrent mutation pattern? Assuming that the AAM process in our mice followed a single-step selection scheme (**Figure 1A**), we can make a first rough estimation. Consider the probability p_k^L to obtain a particular pattern of k mutations out of all possible patterns of k mutations, which is $p_k^L = \frac{1}{M_k^L}$, with $M_k^L = 3^k \times \binom{L}{k}$ being the number of possible mutation patterns of size k, and L being the V sequence length. When an Ab V gene of length $L = 300$ and k mutations is produced by the SHM process, the probability that the outcome is a particular mutation pattern is $p_1^{300} \approx 10^{-3}$ for $k = 1$, and $p_2^{300} \approx 10^{-6}$ for $k = 2$. During an AAM process, thousands of mutated B cells are generated in a mouse; hence, the probability of finding a given mutation pattern of size k among all mutated B cells is $1 - (1 - p_k^{300})^N$, where N is the number of B cells with k mutations. Let us assume that $N = 10^5$ B cells got $k = 2$ mutations. Then, as a crude estimation, the probability of a given mutation pattern of size $k = 1$ and $k = 2$ among all those B cells is, respectively, $1 - (1 - p_1^{300})^{10^5} \approx 1$ and $1 - (1 - p_2^{300})^{10^5} \approx 0.1$. This means that recurrent mutation patterns of such a small size are rather likely to appear in a single-step setting, and are therefore not suitable to contradict the null-hypothesis.

However, for $k = 5$, the probability of obtaining a particular mutation pattern by chance among 10^5 B cells is only $1 - (1 - p_5^{300})^{10^5} \approx 2 \times 10^{-8}$. Hence it is highly improbable that both mice in our thought experiment could produce, by a single-step process, the same recurrent mutation pattern. On the other hand, in a multiple-step selection process, single mutations can be selected one by one (**Figure 1B**). Therefore, finding recurrent mutation patterns of that size or larger would be consistent with a multiple-step scheme while deeming the

single-step null-hypothesis highly improbable. Admittedly, the above is a simplified probability calculation. A more realistic estimation, based on calculations that include reversions and different baseline mutabilities, does not change the above conclusions (see Supplementary Material).

Data from experiments along the main idea of our thought experiment do exist. For instance, Rag1$^{-/-}$ double transgenic mice for Ab *H* and *L* chains are available (25). Also, hapten-conjugated proteins can yield a large percentage of canonical *V* gene sequences (3).

In a survey, we found a number of publications that present Ab *V* sequences obtained from syngeneic mice under the same immunization protocol (3, 4, 26–31). In all the data analyzed so far we could not find a single instance of mutated *V* sequences from GC B cells sharing three or more mutations. Also, a substantial set of independent murine V_H genes with a common V_H germline sequence was recently collected from literature and examined for recurrent patterns (32). The search yielded not a single case of shared triplets.

In summary, to our knowledge, there is no published independent sequence data that contradicts the single-step hypothesis.

CHALLENGE FOR FUTURE RESEARCH: TRYING TO FALSIFY THE SINGLE-STEP HYPOTHESIS WITH HIGH-THROUGHPUT SEQUENCE DATA

A clear understanding of AAM requires answering the question whether the single-step or the multiple-step selection hypothesis hold. A straightforward approach would be direct observation of SHM and Ab-mediated selection events via *in vivo* imaging, but this is technically not yet possible. Similarly daunting is to try to infer the frequency of selection steps from indirect observations while making use of non-validated assumptions.

Our proposal of falsifying the single-step null-hypothesis provides a way out. This hypothesis does not preclude the knowledge of any mechanisms besides the stochastic process of SHM. Moreover, this knowledge does not need to be highly precise because an *upper* estimation of probabilities under the null-hypothesis can suffice.

Therefore, examining Ab *V* gene sequence data with the aim of falsifying the single-step hypothesis is a powerful technique.

Next generation sequencing currently allows to obtain suitable Ab *V* gene sequence data (33, 34). One strategy for the search of contradictions is to calculate, under the null-hypothesis, the probability distribution of recurrent mutation patterns acquired independently.

A detailed calculation of probability distributions is shown in the Supplementary Material. It allows to estimate that, for given realistic parameters (see Table therein), the probabilities of observing recurrent patterns are much lower than 0.05. In case of a very strict multiple-step selection, the null-hypothesis can potentially be contradicted with very few sequences.

This strategy can be pursued both for independent and same-lineage *V* gene sequence sets. In the latter case, the probability calculation must be performed exclusively for recurrent patterns that cannot possibly stem from common ancestors.

Another strategy comprises trying to contradict the null-hypothesis by examining the structure of a same-lineage *V* sequence population for signs of a directed multi-step process as in contrast to an undirected, random process. Such signs can be, for instance, the emergence of independent quasi-species (35), or of coalescence times typical to multi-step processes (36).

These methods require however: (i) that the AAM process has been ongoing long enough for population structures to have emerged, and (ii) that enough sequences can be retrieved to make these structures visible. It is well possible that times are too short and clonal sizes too small to provide this sort of data.

No matter which strategy turns out to be the best, important challenges are still open. For instance, present methods of calculating the pairwise probability that sequences pertain to a common or to a different B cell lineage (32) need to be improved, especially where short junctional regions make identification of lineage difficult. With such an analysis working, different independent Ab *V* sequence sets can also be retrieved from the same individual. A further challenge consists of devising estimators of the recurrent mutation pattern probability distributions adequate to the respective experimental setup. Good estimations of baseline mutability would be helpful; however, using upper estimations of probability might be sufficient.

For pinning down the actual AAM process, it is not advisable to examine data sets that include sequences of memory cells, to avoid the risk of analyzing repeated rounds of immunizations against the same or different antigens. Thus, the design of experiments that consider both the anatomical compartment from which B cells are taken and strategies that maximize the size of data sets, poses a challenge as well.

While multiple-step selection points to AAM as an accelerated molecular evolution process maximizing Ab affinity increase, single-step selection points at an optimization process of Ab repertoires in which both Ab affinity enhancement and diversification can be equally relevant (14, 17, 37). Striving to discover which is right must be a priority to those interested in unveiling AAM mechanisms. Trying to falsify the single-step hypothesis is not easy and might be even impossible – for instance, if the underlying process is indeed a single-step one. But it is, in our opinion, the only viable way.

ACKNOWLEDGMENTS

The authors wish to acknowledge Emilio Faro (Department of Mathematics II, University of Vigo) for his assistance with the calculations in the Supplementary Material. This work was partially supported by the European Union 7th Framework Programme (FP7/REGPOT-2012-2013.1, EC) under grant agreement no. 316265, BIOCAPS. Jose Faro acknowledges the support of PIRSES-GA-2012-317893 (7th FP, EC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00237/full

REFERENCES

1. Eisen HN, Siskind GW. Variations in affinities of antibodies during the immune response. *Biochemistry* (1964) **3**:996–1008. doi:10.1021/bi00895a027
2. Flajnik MF. Comparative analyses of immunoglobulin genes: surprises and portents. *Nat Rev Immunol* (2002) **2**(9):688–98. doi:10.1038/nri889
3. Berek C, Berger A, Apel M. Maturation of the immune response in germinal centers. *Cell*

(1991) **67**(6):1121–9. doi:10.1016/0092-8674(91)90289-B
4. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. *Nature* (1991) **354**(6352):389–92. doi:10.1038/354389a0
5. Griffiths GM, Berek C, Kaartinen M, Milstein C. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature* (1984) **312**(5991):271–5. doi:10.1038/312271a0
6. Berek C, Milstein C. Mutation drift and repertoire shift in the maturation of the immune-response. *Immunol Rev* (1987) **96**:23–41. doi:10.1111/j.1600-065X.1987.tb00507.x
7. MacLennan IC, Gray D. Antigen-driven selection of virgin and memory B cells. *Immunol Rev* (1986) **91**:61–85. doi:10.1111/j.1600-065X.1986.tb01484.x
8. Agur Z, Mazor G, Meilijson I. Maturation of the humoral immune response as an optimization problem. *Proc Biol Sci* (1991) **245**(1313):147–50. doi:10.1098/rspb.1991.0101
9. MacLennan IC, Johnson GD, Liu YJ, Gordon J. The heterogeneity of follicular reactions. *Res Immunol* (1991) **142**(3):253–7. doi:10.1016/0923-2494(91)90070-Y
10. Kepler TB, Perelson AS. Somatic hypermutation in B cells: an optimal control treatment. *J Theor Biol* (1993) **164**(1):37–64. doi:10.1006/jtbi.1993.1139
11. Moreira JS, Faro J. Re-evaluating the recycling hypothesis in the germinal centre. *Immunol Cell Biol* (2006) **84**(4):404–10. doi:10.1111/j.1440-1711.2006.01443.x
12. Raoof S, Heo M, Shakhnovich EI. A one-shot germinal center model under protein structural stability constraints. *Phys Biol* (2013) **10**(2):025001. doi:10.1088/1478-3975/10/2/025001
13. Manser T. Textbook germinal centers? *J Immunol* (2004) **172**(6):3369–75. doi:10.4049/jimmunol.172.6.3369
14. Longo NS, Lipsky PE. Why do B cells mutate their immunoglobulin receptors? *Trends Immunol* (2006) **27**(8):374–80. doi:10.1016/j.it.2006.06.007
15. Or-Guil M, Wittenbrink N, Weiser AA, Schuchhardt J. Recirculation of germinal center B cells: a multilevel selection strategy for antibody maturation. *Immunol Rev* (2007) **216**:130–41. doi:10.1111/j.1600-065X.2007.00507.x
16. Allen CD, Okada T, Cyster JG. Germinal-center organization and cellular dynamics. *Immunity* (2007) **27**(2):190–202. doi:10.1016/j.immuni.2007.07.009
17. Faro J, Combadao J, Gordo I. Did germinal centers evolve under differential effects of diversity vs affinity? In: Bersini H, Carneiro J, editors. *Artificial Immune Systems: 5th International Conference, ICARIS 2006, Oeiras, Portugal, September 4-6, 2006: Proceedings. Lecture Notes in Computer Science*. New York, NY: Springer (2006). p. 1–8.
18. Faro J, Or-Guil M. Reassessing germinal centre reaction concepts. In: Molina-París C, Lythe G, editors. *Mathematical Models and Immune Cell Biology*. New York, NY: Springer (2011). p. 241–58.
19. Faro J, Or-Guil M. How oligoclonal are germinal centers? A new method for estimating clonal diversity from immunohistological sections. *BMC Bioinformatics* (2013) **14**(Suppl 6):S8. doi:10.1186/1471-2105-14-S6-S8
20. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) **338**(1–2):67–74. doi:10.1016/j.jim.2008.06.006
21. Dunn-Walters DK, Belelovsky A, Edelman H, Banerjee M, Mehr R. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Dev Immunol* (2002) **9**(4):233–43. doi:10.1080/10446670310001593541
22. Shahaf G, Barak M, Zuckerman NS, Swerdlin N, Gorfine M, Mehr R. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: a large-scale simulation study. *J Theor Biol* (2008) **255**(2):210–22. doi:10.1016/j.jtbi.2008.08.005
23. Wittenbrink N, Weber TS, Klein A, Weiser AA, Zuschratter W, Sibila M, et al. Broad volume distributions indicate nonsynchronized growth and suggest sudden collapses of germinal center B cell populations. *J Immunol* (2010) **184**(3):1339–47. doi:10.4049/jimmunol.0901040
24. Wittenbrink N, Klein A, Weiser AA, Schuchhardt J, Or-Guil M. Is there a typical germinal center? A large-scale immunohistological study on the cellular composition of germinal centers during the hapten-carrier-driven primary immune response in mice. *J Immunol* (2011) **187**(12):6185–96. doi:10.4049/jimmunol.1101440
25. Paus D, Phan TG, Chan TD, Gardam S, Basten A, Brink R. Antigen recognition strength regulates the choice between extrafollicular plasma cell and germinal center B cell differentiation. *J Exp Med* (2006) **203**(4):1081–91. doi:10.1084/jem.20060087
26. Kallberg E, Gray D, Leanderson T. Kinetics of somatic mutation in lymph node germinal centres. *Scand J Immunol* (1994) **40**(5):469–80. doi:10.1111/j.1365-3083.1994.tb03492.x
27. Ziegner M, Steinhauser G, Berek C. Development of antibody diversity in single germinal centers: selective expansion of high-affinity variants. *Eur J Immunol* (1994) **24**(10):2393–400. doi:10.1002/eji.1830241020
28. Kimoto H, Nagaoka H, Adachi Y, Mizuochi T, Azuma T, Yagi T, et al. Accumulation of somatic hypermutation and antigen-driven selection in rapidly cycling surface Ig+ germinal center (GC) B cells which occupy GC at a high frequency during the primary anti-hapten response in mice. *Eur J Immunol* (1997) **27**(1):268–79. doi:10.1002/eji.1830270140
29. Jacob J, Przylepa J, Miller C, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. III. The kinetics of V region mutation and selection in germinal center B cells. *J Exp Med* (1993) **178**(4):1293–307. doi:10.1084/jem.178.4.1293
30. Han S, Zheng B, Dal Porto J, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. IV. Affinity-dependent, antigen-driven B cell apoptosis in germinal centers as a mechanism for maintaining self-tolerance. *J Exp Med* (1995) **182**(6):1635–44. doi:10.1084/jem.182.6.1635
31. Vora KA, Tumas-Brundage K, Manser T. Contrasting the in situ behavior of a memory B cell clone during primary and secondary immune responses. *J Immunol* (1999) **163**(8):4315–27.
32. Weiser AA, Wittenbrink N, Zhang L, Schmelzer AI, Valai A, Or-Guil M. Affinity maturation of B cells involves not only a few but a whole spectrum of relevant mutations. *Int Immunol* (2011) **23**(5):345–56. doi:10.1093/intimm/dxr018
33. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* (2012) **135**(3):183–91. doi:10.1111/j.1365-2567.2011.03527.x
34. Liberman G, Benichou J, Tsaban L, Glanville J, Louzoun Y. Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions. *Front Immunol* (2013) **4**:274. doi:10.3389/fimmu.2013.00274
35. Eigen M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* (1971) **58**(10):465–523. doi:10.1007/BF00623322
36. Brunet E, Derrida B. Genealogies in simple models of evolution. *J Stat Mech* (2013) **2013**:P01006. doi:10.1088/1742-5468/2013/01/P01006
37. Baumgarth N. How specific is too specific? B-cell responses to viral infections reveal the importance of breadth over depth. *Immunol Rev* (2013) **255**(1):82–94. doi:10.1111/imr.12094

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 October 2013; accepted: 07 May 2014; published online: 26 May 2014.
*Citation: Or-Guil M and Faro J (2014) A major hindrance in antibody affinity maturation investigation: we never succeeded in falsifying the hypothesis of single-step selection. Front. Immunol. **5**:237. doi: 10.3389/fimmu.2014.00237*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Or-Guil and Faro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 09 August 2013
doi: 10.3389/fimmu.2013.00235

A temporal model of human IgE and IgG antibody function

Andrew M. Collins* and Katherine J. L. Jackson

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Tim L. Manser, Thomas Jefferson University, USA
Ramit Mehr, Bar-Ilan University, Israel
Michal Or-Guil, Humboldt University Berlin, Germany

****Correspondence:***
Andrew M. Collins, School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: a.collins@unsw.edu.au

The diversity of the human antibody repertoire that is generated by V(D)J gene rearrangement is extended by nine constant region genes that give antibodies their complex array of effector functions. The application of high throughput sequencing to the study of V(D)J gene rearrangements has led to significant recent advances in our understanding of the antigen-binding repertoire. In contrast, our understanding of antibody function has changed little, and mystery still surrounds the existence of four distinctive IgG subclasses. Recent observations from murine models and from human studies of VDJ somatic point mutations suggest that the timing of emergence of cells from the germinal center may vary as a consequence of class switching. This should lead to predictable differences in affinity between isotypes. These differences, and varying abilities of the isotypes to fix complement and bind FcRs, could help coordinate the humoral defenses over the time course of a response. We therefore propose a Temporal Model of human IgE and IgG function in which early emergence of IgE sensitizes sentinel mast cells while switching to IgG3 recruits FcγR-mediated functions to the early response. IgG1 then emerges as the major effector of antigen clearance, and subsequently IgG2 competes with IgG1 to produce immune complexes that slow the inflammatory drive. Persisting antigen may finally stimulate high affinity IgG4 that outcompetes other isotypes and can terminate IgG1/FcγR-mediated activation via the inhibitory FcγRIIB. In this way, IgG antibodies of different subclasses, at different concentrations and with sometimes opposing functions deliver cohesive, protective immune function.

Keywords: IgG subclasses, humoral immunity, class switching, affinity maturation, IgE, antibody function, B cell differentiation

It is almost 50 years since the complete set of human antibody isotypes was first described (1). For over 30 years, associations have been explored between antibody classes and subclasses and the response to particular pathogens (2). And for almost 30 years, the relationships between cytokine production and antibody class switching have been reported (3). Other rich sources of data that have guided thinking about antibody isotype function have been studies of immunodeficiencies, and the disease susceptibilities with which they are associated (2, 4). Yet despite literally thousands of such studies, and despite significant insights into the particularities of humoral immunity, no proposal has emerged that describes how IgG antibody subclasses and other antibody isotypes *work together* to provide protective immune functions. Here we propose a Temporal Model of human IgE and IgG antibody function, in which there is a programed order to the emergence of the different IgG isotypes that reflects their genomic organization, with switching and emergence being promoted or delayed at different critical points through the action of cytokines. We suggest that early in the germinal center reaction, IgM^+ B cells switch to both IgE and IgG3. Subsequently, IgG1 cells switch and emerge, followed by IgG2-committed cells and finally, if antigen persists, by IgG4-producing cells.

The Temporal Model has its genesis in recent observations of IgE-switched cells in the mouse. These studies suggest that the IgE response is not usually a late development arising from an expanded clone of IgG-committed cells that develops through the germinal center reaction. Rather, it has been shown that IgE class-switched murine cells usually develop and exit the germinal center reaction in the early phase of an immune response, and that they rapidly differentiate into plasmablasts and plasma cells (5, 6). The IgE-secreting plasma cells carry fewer somatic point mutations in their rearranged V(D)J genes than IgG-secreting plasma cells (6), and as a consequence their secreted antibodies are likely to be of lower affinity.

There can be no doubt that IgE antibodies can also be produced late in a response. Recent studies have confirmed the existence of high affinity IgE, and of sequential switching to IgE within the germinal centers of mice (7, 8). No attempt has been made here to incorporate such late IgE into the model. The functions of secretory IgA in mucosal secretions and of serum IgA are also not considered, but the temporal model provides a coherent view of the separate and joint activities of early IgE and the IgG subclasses.

Reports of early IgE in murine models provide a new perspective from which to consider some unusual features of human IgE antibody gene sequences. We have shown that IgE-associated VDJ genes from non-allergic individuals carry very few somatic point mutations, and some IgE sequences carry no mutations at all (9). In individuals with atopic dermatitis, unmutated sequences have also been seen at relatively high frequency (10). In parasitized individuals, we have seen more highly mutated IgE sequences (11),

but these sequences did not carry the pattern of mutations that is considered the mutational signature of antigen selection within the germinal center reaction (12). In some, though not all allergic conditions, IgE sequences also lack this pattern of mutation (9, 10).

These studies can be understood if IgE class switching in humans, as in the mouse, can occur early in the germinal center reaction, and if such switching is rapidly followed by the differentiation of IgE-switched cells into plasmablasts that leave the germinal centers. Some continuing accumulation of somatic point mutations might then take place, outside the germinal centers (13). This would give the mutations in those IgE sequences a distinctly different pattern to that which is seen in IgG sequences that emerge after multiple rounds of selection within the germinal centers. Such selection typically leads to an accumulation of nonsynonymous (replacement) mutations in the complementarity determining regions of the antibody genes (12).

In the context of invasion by pathogens, the production of early IgE antibodies could allow widely dispersed mast cells to function as sentinel cells (14), alerting the immune system to further incursions or spread of the pathogens. Early IgE could function in this way, despite its low affinity, because low affinity IgE has been shown to function well on the surface of mast cells and basophils, if it is directed against multiple epitopes on multivalent antigen (15, 16).

If class switching to IgE is rapidly followed by departure of cells from the germinal center, the possibility that switching to other isotypes may lead cells to follow other distinct developmental pathways cannot be ignored. We have therefore reconsidered the functions of human IgG subclasses, and this has been done in the light of our observations of somatic point mutations in antibodies of different IgG subclasses. These observations provide the broadest possible overview of humoral immunity. In an analysis of almost 1,000 VDJ genes isolated from people living in an area of endemic parasitism, a surprising and statistically significant relationship was seen (11). IgG3-associated VDJ genes were the least mutated VDJ gene sequences, and the mean number of mutations seen in sequences associated with the other subclasses corresponded to the position of each constant region gene within the IGH gene locus. That is, IgG3 < IgG1 < IgG2 < IgG4.

We hypothesize that differences in mean levels of mutation arise because human B cells tend to follow a programed sequence of class switching and departure from the germinal center reaction. We propose that cells first switch from IgM to IgG3, then to IgG1 and to IgG2 and finally to IgG4 following the genomic ordering of the constant region genes (**Figure 1**). This is not to deny the reality of alternative switch pathways under the influence of particular cytokines (17). We propose that class switching is driven by underlying probabilities, and switching is linked to emergence from the germinal centers, leading to the generalizable sequence of the Temporal Model. Through changes in probabilities associated with the expression of adhesion molecules and chemokine receptors, switching could be closely followed by emergence, or emergence could follow variable periods of proliferation, mutation, and selection within the germinal centers. The model does not attempt to resolve the timing of these events for each isotype.

The Temporal Model has parallels with models of division-linked phenotypic change, including class switching, which suggest that predictable order can emerge from stochastic processes

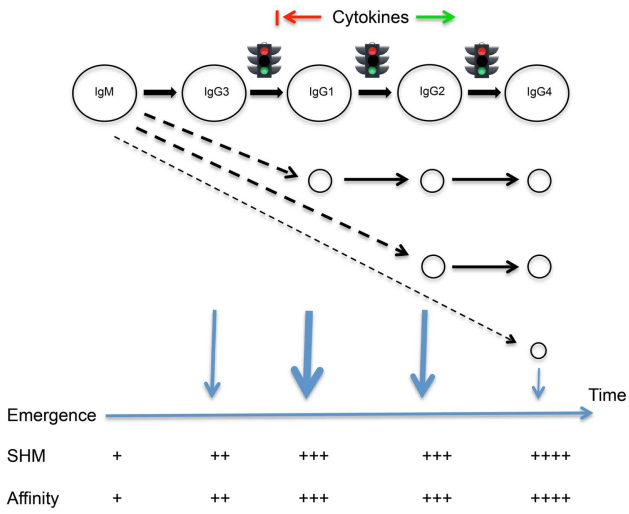


FIGURE 1 | The Temporal Model of IgE/IgG class switching and departure from the germinal center reaction. A programed sequence of sequential switching is highlighted, though alternative switch pathways are also indicated. The timing of switching events and the emergence of different cell types from the germinal center reaction is suggested by the extent of somatic point mutations carried by the cells. Many cytokines may act to promote or delay class switching, as indicated.

because of differences in the underlying probabilities of different outcomes (18, 19). It also is in line with modeling of the dynamics of murine division-linked isotype switching that suggested that the outcome of isotype switching, under the indirect influence of cytokines, is biased toward switching to the immediate downstream neighboring constant region gene (20).

Though a simple relationship between mutation numbers and affinity in any sequence cannot be assumed, accumulating mutations are generally considered to give rise to higher affinity antibodies through selection within the germinal centers (21). Sequential departure of cells from the germinal centers should therefore ensure that antibodies of different isotypes have predictable differences in affinity. This in turn should ensure that despite the different isotypes having some opposing actions, and despite the changing relative concentrations of the different isotypes over time (22), all antibodies at the time of their production should be able to play their assigned roles. It should also ensure that inflammatory processes are tightly controlled, through the temporal coordination of antibodies that have striking differences in their abilities to bind FcγR and to fix complement.

There is good evidence that the IgG3 response occurs early, is relatively transient and is of relatively low affinity (22, 23). This is supported by our sequencing study, for IgG3-associated VDJ genes had the fewest mutations of the IgG subclasses (mean 17.7 mutations), 31% of the sequences had less than 10 mutations and 7% of the sequences had no mutations at all (11). We propose that class switching to IgG3, the first IgG subclass gene in the human IGHC locus, first brings beneficial FcγR-mediated defenses into play. The accumulation of some somatic point mutations during the differentiation of IgG3-committed cells should ensure that

most IgG3 antibodies have experienced some affinity maturation, and the specific physicochemical properties of IgG3 should mean that the switch from IgM to IgG3 does not lead to a crippling loss of binding avidity.

The principal "early" antibody, IgM, is able to provide useful protection despite its low affinity, because of the multivalent nature of secreted IgM, and because of its flexibility (24). The long hinge region of IgG3 makes it the most flexible human IgG antibody (25). This should facilitate bivalent binding of high avidity to repeated determinants on the surface of an invading pathogen. As part of the early response, IgG3 antibodies would have to work with IgM antibodies to efficiently trigger complement fixation and engagement with FcγR-bearing cells. In fact, IgG3 has the highest affinity of the IgG subclasses for C1q, the first component of the classical complement cascade (26). It also has the highest affinity for the FcγRIIIA and FcγRIIIB receptors, and its affinity for FcγRIIA is second only to IgG1 (27).

The elongated hinge region makes IgG3 vulnerable to catabolism. IgG3 has a half-life of just 7 days (28) and shares a short half-life with IgM (∼5 days) (29) and IgE (∼3 days) (30). This rapid turnover of all three kinds of "early antibody" should facilitate the ever-increasing dominance, as a response progresses, of higher affinity antibodies of other isotypes.

In our study of VDJ rearrangements, IgG1-associated sequences were significantly more mutated than IgG3 sequences. The mean mutation of VDJ utilizing the IgG1 gene, positioned immediately downstream from IgG3, was 21.0 somatic point mutations, and only 13% of sequences had fewer than 10 mutations (11). We therefore suggest that IgG1-committed cells are the next cell type to differentiate and depart the germinal centers. Although having on average just three more mutations than IgG3 sequences, we suggest that a number of days are likely to separate the average time of departure of IgG3-committed cells and IgG1-committed cells. It is generally accepted that mutations accumulate at the rate of about one mutation per cell division (31), and centroblast division time is thought to be around 7 h (32). It is likely, however, that as increasing numbers of mutations accumulate, the probability that further random mutations are beneficial is low (33). The speed with which selected sequences accumulate mutations is therefore likely to slow over the course of a response.

Class switching to IgG1 leads to the secretion of more highly mutated, complement-fixing, FcγR-binding IgG antibodies that often dominate the response to bacterial and viral invaders (34, 35). Certainly, IgG1 antibodies are the most abundant serum antibodies (36). With their shorter hinge regions, IgG1 molecules lack flexibility but with their higher affinity for antigen, even monovalent binding should be stable and effective. And with their high affinity for C1q (26) and FcγRI, FcγRII, and FcγRIII (27), such IgG1 antibodies would continue driving inflammatory processes and antigen clearance.

Many studies report that IgG1 antibodies appear relatively early in the immune response, and in fact IgG1 and IgG3 are often the only IgG subclasses detected in a response (37, 38). This could result from early antigen clearance preventing the appearance of IgG2 and IgG4 antibodies. It could also reflect delays in downstream class switching as a result of the prevailing cytokine milieu. T cell cytokines are often said to drive class switching to IgG1 and IgG3 (39). Alternatively, they could be said to delay class switching to IgG2 and IgG4, by their promotion of the IgG1 response.

IgG2 antibodies are the second most abundant serum antibodies and the IgG2 gene is positioned immediately downstream of IgG1. IgG2 antibodies are seen at concentrations that are comparable to IgG1 antibodies, and that are much higher than the typical serum concentrations of IgG3 and IgG4. In contrast to IgG1 antibodies, IgG2 antibodies fix complement very poorly (40, 41) and interact very weakly with FcγR (27). In our sequence study, IgG2 antibodies carried a higher mean number of mutations (22.0) than IgG1 antibodies (21.0) (11). It is difficult to believe that this higher level of mean mutations would lead to biologically significant differences in mean antibody affinity, but certainly IgG2 antibodies must share high affinity with IgG1 antibodies. All these features of the humoral response require explanation. In particular, any model of IgG subclass function must explain how IgG1 and IgG2 antibodies, the two most abundant antibody isotypes, work together to deliver protective immunity despite their diametrically opposed properties.

We hypothesize that IgG2-committed cells emerge from the germinal center reaction, on average, shortly after the development of the IgG1 response. We further hypothesize that IgG2 functions as an anti-inflammatory "partner" to more inflammatory IgG1 antibodies, "dampening down" the inflammatory response by its competition with the IgG1 isotype (**Figure 2**). Working together, IgG1 and IgG2 antibodies could provide a spectrum of activity, from the highly inflammatory "pure IgG1 response," to the "pure IgG2 response" that results in immune complexes that cannot interact with FcγR-bearing cells or with molecules of the complement system.

Mutation data suggests that IgG1 and IgG2 antibodies have similar affinities. If this is the case, they will compete on a level playing field, where antibody concentrations prevail. We propose that the relative concentrations of IgG1 and IgG2 are the result of the balance of cytokines that either promote or delay switching to IgG2. These proportions will be seen in the immune complexes that form during the response, and the outcome of varying proportions of IgG1 and IgG2 antibodies will be immune complexes having varying avidity for complement and for FcγR. IgG2 can therefore be conceptualized as an anti-inflammatory brake on the inflammatory actions of IgG1. In certain circumstances, switching may occur quickly, leading to a response that is dominated by IgG2. This could help explain reports that IgG2 dominates the antibody

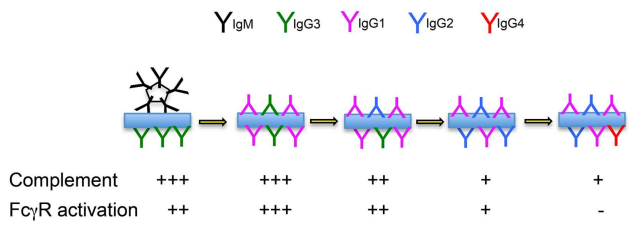


FIGURE 2 | Immune complex formation over the course of a prolonged immune response. The likely complement-fixing and FcγR-binding abilities of immune complexes formed by antibodies of mixed isotypes are indicated.

response to carbohydrates response (42). In fact similar concentrations of IgG1 and IgG2 antibodies have often been reported in response to carbohydrate antigens (43–45), and IgG2 antibodies are also a conspicuous part of the response to many protein antigens (46). It is clear that the chemistry of carbohydrate antigens cannot explain the IgG2 response.

The human IgG4 response is often described as an anti-inflammatory blocking response (47), but the apparent functions of these antibodies have been difficult to reconcile with their very low concentrations. We believe that their mode of action is revealed by the levels of mutations that are seen in IgG4-associated VDJ sequences. IgG4 is the most distal of the IgG subclass genes within the heavy chain constant region locus. In our sequence study, IgG4 antibodies carried the highest mean number of VDJ mutations of the IgG subclasses (mean: 27.1), and no unmutated IgG4-associated VDJ sequences were seen (11). This suggests to us that IgG4-committed cells are (typically) the last cell type to emerge from the germinal center reaction. They are therefore likely to be the highest affinity antibodies. This is indirectly supported by the circumstances in which IgG4 antibodies are conspicuous. Serum IgG4 concentrations are elevated in chronic helminth and other parasite infections (47). Serum IgG4 concentrations also rise during allergy desensitization therapy, after repeated exposure to low doses of allergen (47). They have also been reported in the convalescent phase of the anti-viral response (35).

IgG4 antibodies do not fix complement and bind very poorly to activating FcγR (27), but they bind to the inhibitory FcγRIIB with an affinity that is higher than that of the other three IgG isotypes (27). Critical to the blocking activity of IgG4, inhibition is only mediated via the FcγRIIB receptor when immune complexes co-engage FcγRIIB and other activating FcRs (48). Despite high concentrations of specific antibodies of other isotypes, IgG4 should therefore block FcγR-mediated processes if it is present as a modest proportion of all antibodies in an immune complex. The high affinity of IgG4 should provide it with a competitive advantage, ensuring its participation in immune complex formation, and therefore allowing it to successfully act through FcγRIIB.

The ability of IgG4 antibodies to outcompete other isotypes may also be facilitated by the phenomenon of Fab arm exchange. In reducing conditions, IgG4 antibodies have the unique ability to dissociate into monovalent heavy/light chain pairs, and to re-associate again as bivalent antibodies (49, 50). This Fab arm exchange leads frequently to the formation of bi-specific antibodies. It has been suggested that this would, in practice, lead IgG4 antibodies to be functionally monovalent, as Fab arm exchange would be unlikely between antibodies of related specificities (49). We believe an alternative explanation of the consequences of Fab arm exchange could be the formation of blocking antibodies that bind with very high avidity because of their bi-specific nature.

Bivalency gives power to the IgG molecule. It most obviously allows a single antibody molecule to aggregate two antigen molecules, but it also allows high avidity binding to suitably spaced, repeated epitopes on the surface of a complex antigen. An additional outcome of bivalency has also been identified. It was first proposed on theoretical grounds that very weak binding of one arm of a bivalent antibody molecule to a "non-target" epitope could substantially improve the avidity of binding of the antibody to its target epitope, and that the probability of such bi-specific interactions was reasonably high (51). Recently, such bi-specific heteroligation was shown to facilitate antibody binding to HIV-gp140 (52). For a number of antibodies, high affinity interactions between gp140 and one antigen-binding site were supported by low affinity binding of the second antibody arm to completely different epitopes (52).

Fab arm exchange by IgG4 antibodies could improve the likelihood of heteroligation, if exchange occurs between antibodies of related specificities. This would be likely in two situations. IgG4 antibodies are undetectable in response to many antigens, and individuals with very low serum IgG4 concentrations are likely to have a limited IgG4 repertoire. In such circumstances, Fab arm exchange between antibodies targeting associated epitopes or related antigens would be more likely. In individuals with higher IgG4 concentrations, Fab arm exchange could also take place between antibodies of related specificities through their co-localization at sites of inflammation. At such sites, appropriate redox conditions of even a transient nature could lead Fab arm exchange to "lock together" Fab arms of associated specificities. This would function to increase IgG4 binding avidity, giving bi-specific IgG4 antibodies the ability to outcompete the more inflammatory isotypes, late in a response.

In addition to its IgG1-blocking activity, it is clear that IgG4 can block IgE-mediated immune function. The IgE and IgG4 isotypes are strongly linked with one another in the literature, and in fact IgE antibodies are often said to arise by class switching of $IgG4^+$ B cells (53, 54). The mutational characteristics that we have reported clearly demonstrate that this is not always so, for both the number and patterns of mutations we have seen differ between IgE and IgG4 (11). However it is possible that IgG4 antibodies could be related to high affinity IgE, if such IgE antibodies arise late in a response. Clarification of this question, and determination of the circumstances in which such high affinity IgE might be produced by the human immune system will be necessary before late-arising IgE can be incorporated into the model.

The temporal model, as presented here, outlines sequential class switching during a first, persisting exposure to antigen. The nature of isotype expression in a recall response will clearly depend upon the tendency of class-switched cells to differentiate into memory cells during the primary response. Though IgM memory cells are now known to be an important part of the memory compartment (55, 56), there is some evidence from the early literature that IgG1 dominates the switched memory compartment. Studies of the recall response after re-challenge with keyhole limpet hemocyanin (KLH) showed little or no increase in peak concentrations of KLH-specific IgG2 and IgG3 antibodies, but a marked increase in circulating IgG1 anti-KLH antibodies (57). IgG4 antibodies were only seen upon re-challenge. The high affinity of IgG1 and its inflammatory functions make it the ideal isotype for IgG memory, and the logic of the Temporal Model suggests that memory cells of the other IgG isotypes could be either unhelpful or counterproductive.

The contributions that memory cells make to antibody isotype production in a recall response will depend upon whether memory cells re-enter germinal centers or immediately differentiate into plasmablasts and plasma cells. Studies in the mouse suggest that in a recall response, mouse IgG^+ memory cells (55) and IgE^+ memory cells (8) rapidly give rise to plasma cells, but IgM^+

memory cells re-enter the germinal center reaction (55). If human and mouse cells are governed by similar processes, it may therefore be that events within the human germinal center during a recall response proceed as we have outlined for earlier events. In other words, reactivated IgM memory cells within the germinal centers would give rise to new IgE-switched and IgG-switched cells through a programed process of sequential switching.

Mechanisms that could underlie the temporal emergence of different IgG subclasses from the GC reaction will need to be explored, and one possibility lies in the recently reported competitive feedback between soluble antibody from plasma cells and the GC B cells (58). Soluble antibody, produced by cells that have previously emerged and differentiated, competes with GC B cells for binding to FDC-associated antigen. This competition promotes survival of GC B cells with higher affinity than the soluble antibody, while B cells of lesser affinity die by neglect. The Temporal Model suggests that as class switching proceeds, antibodies expressing subclasses from the more distally positioned IgG genes are likely to be of higher affinity. At a particular point in the response, higher affinity antibodies that express downstream IgG genes would outcompete soluble antibody of earlier subclasses, while promoting the destruction of B cells expressing earlier subclasses that carry fewer mutations and are of lower affinity. Feedback competition may therefore promote the temporal emergence of the subclasses in their genomic order.

Many studies give credence to the Temporal Model, but certainly this is not true of all studies. Discordant observations could be the result of pathogen-directed perturbations of normal immune function, for the temporal progression of isotype switching would be as susceptible as other aspects of immune function to subversion by bacterial and viral virulence factors (59). Discordant observations are particularly seen in some early studies of antibody isotypes, but these reports might be explained by the cross-reactivity of many early "isotype-specific" reagents (60). Others might now be explained by phenomena such as Fc–Fc binding of IgG4 antibodies that were unknown until recently (50). But the resolution of the mystery of antibody function cannot come from studies of the past. It is our hope that this description of the Temporal Model will encourage the question of antibody isotype function to be revived. Having received so little attention over the last two decades, it is now time for the power of high throughput sequencing to be harnessed, to confirm the relationship between the levels of mutation and antibody isotypes in individuals of different ethnicities and states of health, and to properly address the clonal relationships between B cells producing antibodies of different isotypes. It may then be that the timing of class switching, the passage of different cell populations between anatomical compartments within the lymph node, the emergence of cells from the germinal center reaction, and the overall functions of human isotypes can finally be determined with certainty.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Health and Medical Research Council.

REFERENCES

1. Ballieux RE, Bernier GM, Tominaga K, Putnam FW. Gamma globulin antigenic types defined by heavy chain determinants. *Science* (1964) **145**(3628):168–70. doi:10.1126/science.145.3628.168
2. Oxelius VA. Immunoglobulin G (IgG) subclasses and human disease. *Am J Med* (1984) **76**(3A):7–18. doi:10.1016/0002-9343(84)90314-0
3. Snapper CM, Paul WE. Interferon-gamma and B cell stimulatory factor-1 reciprocally regulate Ig isotype production. *Science* (1987) **236**(4804):944–7. doi:10.1126/science.3107127
4. Umetsu DT, Ambrosino DM, Quinti I, Siber GR, Geha RS. Recurrent sinopulmonary infection and impaired antibody response to bacterial capsular polysaccharide antigen in children with selective IgG-subclass deficiency. *N Engl J Med* (1985) **313**(20):1247–51. doi:10.1056/NEJM198511143132002
5. Erazo A, Kutchukhidze N, Leung M, Christ AP, Urban JF Jr., Curotto de Lafaille MA, et al. Unique maturation program of the IgE response in vivo. *Immunity* (2007) **26**(2):191–203. doi:10.1016/j.immuni.2006.12.006
6. Yang Z, Sullivan BM, Allen CD. Fluorescent in vivo detection reveals that IgE(+) B cells are restrained by an intrinsic cell fate predisposition. *Immunity* (2012) **36**(5):857–72. doi:10.1016/j.immuni.2012.02.009
7. Xiong H, Dolpady J, Wabl M, Curotto de Lafaille MA, Lafaille JJ. Sequential class switching is required for the generation of high affinity IgE antibodies. *J Exp Med* (2012) **209**(2):353–64. doi:10.1084/jem.20111941
8. Talay O, Yan D, Brightbill HD, Straney EE, Zhou M, Ladi E, et al. IgE(+) memory B cells and plasma cells generated through a germinal-center pathway. *Nat Immunol* (2012) **13**(4):396–404. doi:10.1038/ni.2256
9. Dahlke I, Nott DJ, Ruhno J, Sewell WA, Collins AM. Antigen selection in the IgE response of allergic and nonallergic individuals. *J Allergy Clin Immunol* (2006) **117**(6):1477–83. doi:10.1016/j.jaci.2005.12.1359
10. Kerzel S, Rogosch T, Struecker B, Maier RF, Zemlin M. IgE transcripts in the circulation of allergic children reflect a classical antigen-driven B cell response and not a superantigen-like activation. *J Immunol* (2010) **185**(4):2253–60. doi:10.4049/jimmunol.0902942
11. Wang Y, Jackson KJ, Chen Z, Gaeta BA, Siba PM, Pomat W, et al. IgE sequences in individuals living in an area of endemic parasitism show little mutational evidence of antigen selection. *Scand J Immunol* (2011) **73**(5):496–504. doi:10.1111/j.1365-3083.2011.02525.x
12. Chang B, Casali P. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunol Today* (1994) **15**(8):367–73. doi:10.1016/0167-5699(94)90175-9
13. Snow RE, Djukanovic R, Stevenson FK. Analysis of immunoglobulin E VH transcripts in a bronchial biopsy of an asthmatic patient confirms bias towards VH5, and indicates local clonal expansion, somatic mutation and isotype switch events. *Immunology* (1999) **98**(4):646–51. doi:10.1046/j.1365-2567.1999.00910.x
14. Marshall JS. Mast-cell responses to pathogens. *Nat Rev Immunol* (2004) **4**(10):787–99. doi:10.1038/nri1460
15. Collins AM, Basil M, Nguyen K, Thelian D. Rat basophil leukaemia (RBL) cells sensitized with low affinity IgE respond to high valency antigen. *Clin Exp Allergy* (1996) **26**(8):964–70. doi:10.1046/j.1365-2222.1996.d01-387.x
16. Gieras A, Focke-Tejkl M, Ball T, Verdino P, Hartl A, Thalhamer J, et al. Molecular determinants of allergen-induced effector cell degranulation. *J Allergy Clin Immunol* (2007) **119**(2):384–90. doi:10.1016/j.jaci.2006.09.034
17. Fujieda S, Zhang K, Saxon A. IL-4 plus CD40 monoclonal antibody induces human B cells gamma subclass-specific isotype switch: switching to gamma 1, gamma 3, and gamma 4, but not gamma 2. *Mian Yi Xue Za Zhi* (1995) **155**(5):2318–28.
18. Tangye SG, Ferguson A, Avery DT, Ma CS, Hodgkin PD. Isotype switching by human B cells is division-associated and regulated by cytokines. *Mian Yi Xue Za Zhi* (2002) **169**(8):4298–306.
19. Tangye SG, Hodgkin PD. Divide and conquer: the importance of cell division in regulating B-cell responses. *Immunology* (2004) **112**(4):509–20. doi:10.1111/j.1365-2567.2004.01950.x
20. Yaish B, Mehr R. Models for the dynamics and order of immunoglobulin isotype switching. *Bull Math Biol* (2005) **67**(1):15–32. doi:10.1016/j.bulm.2004.05.007

21. Shlomchik MJ, Weisel F. Germinal center selection and the development of memory B and plasma cells. *Immunol Rev* (2012) **247**(1):52–63. doi:10.1111/j.1600-065X.2012.01124.x
22. Devey ME, Bleasdale-Barr KM, Bird P, Amlot PL. Antibodies of different human IgG subclasses show distinct patterns of affinity maturation after immunization with keyhole limpet haemocyanin. *Immunology* (1990) **70**(2):168–74. Erratum in: *Immunology* (1990) **71**(1):152.
23. Wilson KM, Di Camillo C, Doughty L, Dax EM. Humoral immune response to primary rubella virus infection. *Clin Vaccine Immunol* (2006) **13**(3):380–6. doi:10.1128/CVI.13.3.380-386.2006
24. Tobita T, Oda M, Azuma T. Segmental flexibility and avidity of IgM in the interaction of polyvalent antigens. *Mol Immunol* (2004) **40**(11):803–11. doi:10.1016/j.molimm.2003.09.011
25. Roux KH, Strelets L, Michaelsen TE. Flexibility of human IgG subclasses. *Mian Yi Xue Za Zhi* (1997) **159**(7):3372–82.
26. Schroeder HW Jr., Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol* (2010) **125**(2 Suppl 2):S41–52. doi:10.1016/j.jaci.2009.09.046
27. Bruhns P. Properties of mouse and human IgG receptors and their contribution to disease models. *Blood* (2012) **119**(24):5640–9. doi:10.1182/blood-2012-01-380121
28. Morell A, Terry WD, Waldmann TA. Metabolic properties of IgG subclasses in man. *J Clin Invest* (1970) **49**(4):673–80. doi:10.1172/JCI106279
29. Barth WF, Wochner RD, Waldmann TA, Fahey JL. Metabolism of human gamma macroglobulins. *J Clin Invest* (1964) **43**:1036–48. doi:10.1172/JCI104987
30. Waldmann TA, Iio A, Ogawa M, McIntyre OR, Strober W. The metabolism of IgE. Studies in normal individuals and in a patient with IgE myeloma. *J Immunol* (1976) **117**(4):1139–44.
31. Sablitzky F, Wildner G, Rajewsky K. Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J* (1985) **4**(2):345–50.
32. Liu YJ, Zhang J, Lane PJ, Chan EY, MacLennan IC. Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur J Immunol* (1991) **21**(12):2951–62. doi:10.1002/eji.1830211209
33. Shannon M, Mehr R. Reconciling repertoire shift with affinity maturation: the role of deleterious mutations. *J Immunol* (1999) **162**(7):3950–6.
34. Biganzoli P, Ferreyra L, Sicilia P, Carabajal C, Frattari S, Littvik A, et al. IgG subclasses and DNA detection of HHV-6 and HHV-7 in healthy individuals. *J Med Virol* (2010) **82**(10):1679–83. doi:10.1002/jmv.21880
35. Deshmukh TM, Shah RR, Gurav YK, Arankalle VA. Serum immunoglobulin G subclass responses in different phases of hepatitis E virus infection. *J Med Virol* (2013) **85**(5):828–32. doi:10.1002/jmv.23537
36. Schauer U, Stemberg F, Rieger CH, Borte M, Schubert S, Riedel F, et al. IgG subclass concentrations in certified reference material 470 and reference values for children and adults determined with the binding site reagents. *Clin Chem* (2003) **49**(11):1924–9. doi:10.1373/clinchem.2003.022350
37. Murphy SL, Li H, Mingozzi F, Sabatino DE, Hui DJ, Edmonson SA, et al. Diverse IgG subclass responses to adeno-associated virus infection and vector administration. *J Med Virol* (2009) **81**(1):65–74. doi:10.1002/jmv.21360
38. Spinsanti LI, Farias AA, Aguilar JJ, del Pilar Diaz M, Contigiani MS. Immunoglobulin G subclasses in antibody responses to St. Louis encephalitis virus infections. *Arch Virol* (2011) **156**(10):1861–4. doi:10.1007/s00705-011-1047-3
39. Pene J, Gauchat JF, Lecart S, Drouet E, Guglielmi P, Boulay V, et al. Cutting edge: IL-21 is a switch factor for the production of IgG1 and IgG3 by human B cells. *J Immunol* (2004) **172**(9):5154–7.
40. Bindon CI, Hale G, Bruggemann M, Waldmann H. Human monoclonal IgG isotypes differ in complement activating function at the level of C4 as well as C1q. *J Exp Med* (1988) **168**(1):127–42. doi:10.1084/jem.168.1.127
41. Bruggemann M, Williams GT, Bindon CI, Clark MR, Walker MR, Jefferis R, et al. Comparison of the effector functions of human immunoglobulins using a matched set of chimeric antibodies. *J Exp Med* (1987) **166**(5):1351–61. doi:10.1084/jem.166.5.1351
42. Barrett DJ, Ayoub EM. IgG2 subclass restriction of antibody to pneumococcal polysaccharides. *Clin Exp Immunol* (1986) **63**(1):127–34.
43. Hammarstrom L, Persson MA, Smith CI. Immunoglobulin subclass distribution of human anti-carbohydrate antibodies: aberrant pattern in IgA-deficient donors. *Immunology* (1985) **54**(4):821–6.
44. Morell A, Doran JE, Skvaril F. Ontogeny of the humoral response to group A streptococcal carbohydrate: class and IgG subclass composition of antibodies in children. *Eur J Immunol* (1990) **20**(7):1513–7. doi:10.1002/eji.1830200716
45. Shackelford PG, Granoff DM, Nelson SJ, Scott MG, Smith DS, Nahm MH. Subclass distribution of human antibodies to *Haemophilus influenzae* type b capsular polysaccharide. *Mian Yi Xue Za Zhi* (1987) **138**(2):587–92.
46. Xu W, Santini PA, Sullivan JS, He B, Shan M, Ball SC, et al. HIV-1 evades virus-specific IgG2 and IgA responses by targeting systemic and intestinal B cells via long-range intercellular conduits. *Nat Immunol* (2009) **10**(9):1008–17. doi:10.1038/ni.1753
47. Aalberse RC, Stapel SO, Schuurman J, Rispens T. Immunoglobulin G4: an odd antibody. *Clin Exp Allergy* (2009) **39**(4):469–77. doi:10.1111/j.1365-2222.2009.03207.x
48. Bruhns P, Fremont S, Daeron M. Regulation of allergy by Fc receptors. *Curr Opin Immunol* (2005) **17**(6):662–9. doi:10.1016/j.coi.2005.09.012
49. van der Neut Kolfschoten M, Schuurman J, Losen M, Bleeker WK, Martinez-Martinez P, Vermeulen E, et al. Anti-inflammatory activity of human IgG4 antibodies by dynamic Fab arm exchange. *Science* (2007) **317**(5844):1554–7. doi:10.1126/science.1144603
50. Rispens T, Ooijevaar-de Heer P, Bende O, Aalberse RC. Mechanism of immunoglobulin G4 Fab-arm exchange. *J Am Chem Soc* (2011) **133**(26):10302–11. doi:10.1021/ja203638y
51. Hodgkin PD. An antigen valence theory to explain the evolution and organization of the humoral immune response. *Immunol Cell Biol* (1997) **75**(6):604–18. doi:10.1038/icb.1997.95
52. Mouquet H, Scheid JF, Zoller MJ, Krogsgaard M, Ott RG, Shukair S, et al. Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature* (2010) **467**(7315):591–5. doi:10.1038/nature09385
53. Jabara HH, Loh R, Ramesh N, Vercelli D, Geha RS. Sequential switching from mu to epsilon via gamma 4 in human B cells stimulated with IL-4 and hydrocortisone. *Mian Yi Xue Za Zhi* (1993) **151**(9):4528–33.
54. Aalberse RC, Platts-Mills TA. How do we avoid developing allergy: modifications of the TH2 response from a B-cell perspective. *J Allergy Clin Immunol* (2004) **113**(5):983–6. doi:10.1016/j.jaci.2004.02.046
55. Dogan I, Bertocci B, Vilmont V, Delbos F, Megret J, Storck S, et al. Multiple layers of B cell memory with different effector functions. *Nat Immunol* (2009) **10**(12):1292–9. doi:10.1038/ni.1814
56. Kurosaki T, Aiba Y, Kometani K, Moriyama S, Takahashi Y. Unique properties of memory B cells of different isotypes. *Immunol Rev* (2010) **237**(1):104–16. doi:10.1111/j.1600-065X.2010.00939.x
57. Bird P, Calvert JE, Amlot PL. Distinctive development of IgG4 subclass antibodies in the primary and secondary responses to keyhole limpet haemocyanin in man. *Immunology* (1990) **69**(3):355–60.
58. Zhang Y, Meyer-Hermann M, George LA, Figge MT, Khan M, Goodall M, et al. Germinal center B cells govern their own fate via antibody feedback. *J Exp Med* (2013) **210**(3):457–64. doi:10.1084/jem.20120150
59. Brodsky IE, Medzhitov R. Targeting of immune signalling networks by bacterial pathogens. *Nat Cell Biol* (2009) **11**(5):521–6. doi:10.1038/ncb0509-521
60. Buckley RH. Immunoglobulin G subclass deficiency: fact or fancy? *Curr Allergy Asthma Rep* (2002) **2**(5):356–60. doi:10.1007/s11882-002-0067-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2013; accepted: 29 July 2013; published online: 09 August 2013.
*Citation: Collins AM and Jackson KJL (2013) A temporal model of human IgE and IgG antibody function. Front. Immunol. **4**:235. doi: 10.3389/fimmu.2013.00235*
This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.
Copyright © 2013 Collins and Jackson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 10 March 2014
doi: 10.3389/fimmu.2014.00086

Self-tolerance in a minimal model of the idiotypic network

Robert Schulz, Benjamin Werner*† and Ulrich Behn*

Institute for Theoretical Physics, University of Leipzig, Leipzig, Germany

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Gennady Bocharov, Russian Academy of Sciences, Russia
Véronique Thomas-Vaslin, Centre National de la Recherche Scientifique, France

****Correspondence:***
Benjamin Werner, Max Planck Institute for Evolutionary Biology, August-Thienemann-Street, 2, D-24306 Plön, Germany
e-mail: werner@evolbio.mpg.de; Ulrich Behn, Institute for Theoretical Physics, University of Leipzig, POB 100 900, D-04009 Leipzig, Germany
e-mail: behn@itp.uni-leipzig.de

†Present address:
Benjamin Werner, Max Planck Institute for Evolutionary Biology, Plön, Germany

We consider the problem of self-tolerance in the frame of a minimalistic model of the idiotypic network. A node of this network represents a population of B-lymphocytes of the same idiotype, which is encoded by a bit string. The links of the network connect nodes with (nearly) complementary strings. The population of a node survives if the number of occupied neighbors is not too small and not too large. There is an influx of lymphocytes with random idiotype from the bone marrow. Previous investigations have shown that this system evolves toward highly organized architectures, where the nodes can be classified into groups according to their statistical properties. The building principles of these architectures can be analytically described and the statistical results of simulations agree very well with results of a modular mean-field theory. In this paper, we present simulation results for the case that one or several nodes, playing the role of self, are permanently occupied. These self nodes influence their linked neighbors, the autoreactive clones, but are themselves not affected by idiotypic interactions. We observe that the group structure of the architecture is very similar to the case without self antigen, but organized such that the neighbors of the self are only weakly occupied, thus providing self-tolerance. We also treat this situation in mean-field theory, which give results in good agreement with data from simulation. The model supports the view that autoreactive clones, which naturally occur also in healthy organisms are controlled by anti-idiotypic interactions, and could be helpful to understand network aspects of autoimmune disorders.

Keywords: idiotypic network, self-tolerance, control of autoreactive idiotypes, autoimmunity, bitstring model, mean-field theory

1. INTRODUCTION

B-lymphocytes express Y-shaped receptor molecules, antibodies, on their surface. These antibodies have specific binding sites which determine their idiotype. All receptors of a given B-cell have the same idiotype. B-cells with random idiotypes of remarkable diversity are produced in the bone marrow.

A B-cell is stimulated to proliferate if its receptors are cross-linked by complementary structures, unstimulated B-cells die. Proliferation occurs if the concentration of complementary structures is not too low or not too high, see e.g., Ref. (1). The latter condition refers to a steric hindrance for cross-linking if too many complementary molecules are around. Stimulating complementary structures can be found on foreign antigens and on other, so-called anti-idiotypic antibodies of complementary specificity. Thus B-lymphocytes can stimulate each other and form a functional network, the idiotypic network, as first proposed in Ref. (2), see also Ref. (3, 4).

The potential repertoire includes idiotypes that can recognize other complementary structures, e.g., on the active sites of enzymes, hormones, and neurotransmitters. Further, there are idiotypic interactions of B-lymphocytes with T-lymphocytes and between T-cells (5). Thus, the idiotypic network is not an autonomous entity of the adaptive immune system, but is coupled to many other networks.

Even for a hypothetical autonomous B-lymphocyte system, we have the requisites of evolution, random innovation, and selection. So the architecture of the idiotypic network can be conceived as the result of an evolution during the life time of an individual. In a revised version of the idiotypic network paradigm, the second generation idiotypic network (6–8), it was suggested that this architecture comprises a densely connected central part with autonomous dynamics and a hereto disconnected (or only sparsely connected) periphery. The periphery is able to clonal expansion in (an adaptive) response to external antigen, and since it is disconnected to the central part, the stimulation does not percolate through the network.

Already Jerne thought the idiotypic network to play an essential role in the control of autoreactive idiotypes (3). Today, the concept of idiotypic networks is still popular in the research on autoimmune diseases, both in theoretical studies and clinical context. Indeed, autoreactive antibodies are regularly found in healthy individuals though in low concentrations. Antibodies which escape other regulatory mechanisms can be controlled by the idiotypic network (9). Anti-idiotypic antibodies specific to potentially autoreactive clones are found in healthy individuals or in patients during remission, they are absent during periods of active autoimmune disease (10). Autoimmune diseases can be related to perturbations of the control of autoreactive clones (10–17), as for example in Myasthenia gravis, a well known B-cell associated autoimmune disease (18).

There are many alternative or complementary concepts to explain self-tolerance and a multitude of possible mechanisms

to cause autoimmune diseases. It is of course beyond the scope of this paper to give an exhaustive review over this rapidly expanding field. We can list only a necessarily subjective selection of a few major concepts and mechanisms. Several theoretical concepts of self-nonself discrimination are presented in a topical issue of Seminars in Immunology (19), including the Two-signal theory (20), the Danger model (21), the context dependent tuning of T-cell antigen recognition (22), cf. also (23, 24), and the Immunological homunculus (25), cf. also (26, 27). Zinkernagel (28) emphasizes the importance of localization, dose, and time of antigens: antigen that does not reach secondary lymphoid organs in minimum doses or for sufficiently long times is immunologically ignored.

Regulatory T-cells have been identified to suppress a variety of immune responses and playing a crucial role in self-tolerance and in controlling the balance of T-helper cells such as Th1, Th2, and Th17 (29, 30). Various mechanisms how infections can trigger autoimmunity are reviewed in Ref. (31). Superantigens may cause a polyclonal T-cell response with an excessive cytokine release, which in turn can induce autoimmune disorders. Chronic tissue damage can, regardless of the initial stimulus, lead to a spreading of the specificity of the T-cell response (epitope spreading) including self-epitopes (32). More recently, epigenetic mechanisms which may cause a breakdown of immune tolerance have been identified in the context of several autoimmune diseases, for a review see Ref. (33), cf. also Ref. (34).

Recent progress in the understanding of autoimmune diseases is reviewed in a topical section of Current Opinion in Immunology edited by Wucherpfennig and Noel (35). The T-cell system and the B-cell system interact in various ways at different stages of an immune response and the distinction between B-cell mediated and T-cell mediated autoimmune disorders appears to erode (36). For T-independent features of B-cell response confer however (37). Also idiotype driven interactions exist between B-cells and T-cells, as reviewed in Ref. (38). Very recently, regulatory B-cells are brought into discussion (36, 39).

There are early attempts to model self-tolerance and autoimmunity mathematically within the network paradigm. We can distinguish papers which consider networks with predefined architecture from work, which studies the (ontogenic) evolution of the networks architecture.

In Ref. (40), based on experimental results (41), an idealized architecture of 26 clones was proposed, which comprises four groups of B-cell clones, a multi-affine group A, two mirror groups B and C with mutual coupling but no intra-group affinity, and a group D which couples with low affinity only to A. Based on this *ad hoc* architecture, a set of non-linear ordinary differential equations (ODEs) is proposed (42) that describes the continuous dynamics of B-cells and antibodies in the presence of self. The proliferation and maturation of by idiotypic interactions activated B-cells is modeled by the non-linear terms of the ODEs. Computer simulations of these ODEs reveal that the response of clones, which couple to self antigen depends on their connectivity to other clones of the network: the higher the connectivity the greater the degree of tolerance; poorly connected clones show unlimited growth.

In Ref. (43), an analytical theory for the dynamics of clones in the mirror groups B and C, which feel the mean-field exerted by the clones of group A that couple to self antigen is considered. The model describes a switching between tolerant and autoimmune states and reverse, induced by infection with external antigen.

Also a paper by Calenbuhr et al. (44) studies the behavior of idiotypic networks with predefined architecture in the presence of self. There, using a similar continuous dynamics as (42) the interaction between N clones of different idiotypes is determined by an $N \times N$ connectivity matrix ($N = 2, \ldots, 25$) with entries zero and one. The maximum number of interactions C of a single clone with other clones is varied between 1 and $N - 1$ and open (chain like) and closed architectures are distinguished. The autonomous system shows oscillatory or chaotic behavior with parameter depending amplitudes. The response to a self-antigen depends on its concentration, and on the parameters of the autonomous system. The state of the system is called tolerant (safe) if the clones which couple to the self have low concentration, otherwise, for a large or even unbounded response, it is called dangerous. The study confirms that more densely connected networks tend to provide tolerant states.

Our work describing the *evolution* of the idiotypic network in the presence of self antigens is similar in spirit to previous work by De Boer and Perelson (45), Stewart and Varela (46), and Takumi and De Boer (47).

De Boer and Perelson (45) investigated a model which describes the population dynamics of antibodies and B-cells by a set of non-linear ODEs. The idiotype is modeled in a discrete shape space by bitstrings of length L ($L = 32$), two idiotypes match if the two aligned bitstrings are complementary in at least T adjacent positions (T is varied from 6 to 11, mainly $T = 8$) which mimics the presence of several idiotopes on an antibody with certain idiotype. For exactly T complementary positions an affinity of 0.1 is assigned, for more than T an affinity of 1. The stimulation of B-cells is described by a bell-shaped activation function, and the production of antibodies by stimulated B-cells by a gearing-up mechanism. There is an input of 10 new clones per day. They are incorporated in the network if at least one other clone is complementary. Clones with too high connectivity are suppressed. Simulations show that the network reaches a stationary regime where the idiotypes that are incorporated in the network are more similar than to be expected for a completely random choice. This gives an advantage because the incorporated B-cells feel a similar stimulating field and their (similar) antibodies do not form complexes. Among the clones which do not expand there are about 25% which have no sufficient stimulation. They are not incorporated in the network and can be considered as the clonal (peripheral) component of the immune system (similar to the singletons in our work, see below). Self antigen is also modeled by bitstrings. In high concentration it suppresses all clones which recognize the antigen, in stimulative concentrations (i.e., if their field is in the stimulating region of the bell-shaped activation function) it gives rise to unlimited self aggression. The authors mention that some of the self-reactive clones, especially those with a high connectivity, are controlled by overstimulation, clones with few connections escape the control.

Stewart and Varela (46) considered a model, which describes the presence or absence of clones of a given idiotype, not distinguishing B-cells and antibodies, using a discrete dynamics. A clone of idiotype i survives if it receives a stimulus σ_i within an allowed window, $\sigma_L \leq \sigma_i \leq \sigma_U$. If σ_i is outside the window, the clone does

not survive the next step of a parallel update. The stimulus an idiotype i receives from clones of complementary idiotype is calculated in a double-sheeted two-dimensional continuous shape space as $\sigma_i = \sum_j m_{ij}$ where $m_{ij} = \exp\{-a_{ij}/c\}^2$. An idiotype is represented by a point on one of the sheets (say, the white one) while the perfectly complementary idiotype has the same coordinates on the other (black) sheet. a_{ij} is the Euklidian distance of two points at different sheets and c is a characteristic distance below which idiotypic interactions are relevant. Simulations for periodic boundary conditions show that stationary patterns on the shape space emerge which consist of nested (concentric) black and white ellipses. They can be conceived as mirror groups where members of one group have only idiotypic interactions with the other group but not within their own group. Idiotypes disconnected from these groups (clonal components) only occur before saturation. The system needs a longer time to reach saturation as smaller c is. Self antigen is represented as points on the two-dimensional shape space. If located on the black (white) sheet it is incorporated in the black (white) elliptic lines. So if the bone marrow is able to produce idiotypes similar to the self, they buffer the self against aggressive autoimmunity.

Takumi and De Boer (47) investigated the evolution of a model network on a double-sheeted two-dimensional discrete shape space in the presence of self-epitopes. Self-reactive clones are deleted by hand assuming some not closer characterized self-tolerance process. Each idiotype has several determinants (idiotopes). New B-cell clones are generated randomly. The dynamics of B-cells is described by a system of ODEs with a log-bell-shaped activation function. A buffering term prohibits the explosion of the clone size, clones are removed if their size falls below an extinction threshold. Their main finding is that the network organizes such that most self-epitopes are embedded in an antibody repertoire of intermediate concentration. Without the explicit deletion of self-reactive clones the authors were unable to obtain robust self-tolerance.

The B-cell models mentioned above, describing the evolution of the network, have in common that they do not show an appropriate partitioning into network and disconnected fraction, and are not reliably stable when coupled to permanently present self antigen. Motivated by these drawbacks (48, 49) proposed to extend their previous models to include the cooperation with T-lymphocytes. Indeed, simulations of the ontogenic evolution of the network in the presence of self antigens ("founder" antigens) show that the system differentiates in several stages into two coexisting compartments, the central immune system that couples to and tolerates self antigens, and the peripheral immune system that could respond to "late" antigen. In the first stage, T-cells which become activated by the initial founding set of antigens, activate in turn B-cells. This continues until the B-cell repertoire is complete and the B-cells start to exert a regulatory feedback on the T-cells. In the second stage, the B-cells compete for T-cell help and their repertoire shrinks to B-cells of an idiotype, which directly recognize a T-cell receptor. After this, a single new antigen would elicit a response only of clones, which are not mounted to the network. However it turned out, that the peripheral system is too tolerant to a later antigen. This motivated (50) to further modify this model making the idiotypic connectivity an explicit function of time, and introducing a log-bell-shaped activation function also for the T-cells. Stewart and Coutinho (51) reviewed the state of modeling and the development of the paradigm, and critically mentioned the lack of experimental evidence supporting the physiological significance of idiotypic interactions between B-cell and T-cell receptors.

For more detailed reviews on the history of the paradigm, mathematical modeling, and new immunological and clinical developments the reader is referred to Ref. (52, 53). For very interesting personal accounts on the development of the network paradigm and the concept of immunological self, see Ref. (8, 54–57).

In the present paper, we consider a model of the idiotypic B-cell network proposed in Ref. (58) which describes the evolution toward complex, functional architectures. The model uses a discrete shape space spanned by bitstrings which represent idiotypes. The discrete dynamics describes presence or absence of idiotypic clones, which survive if their stimulus is within an allowed window. In a sense, the model combines the simplest features of the models previously proposed by De Boer and Perelson (45) and Stewart and Varela (46) and therefore can be considered as a minimal model.

The most interesting architecture emerging in this model comprises (i) densely linked core groups, (ii) peripheral groups without intra-group linking, (iii) groups of suppressed clones, and (iv) groups of singletons which potentially interact only with the suppressed clones. The expressed clones of the core and periphery groups build the actual network, the central part. The expressed clones of the singleton groups are not mounted to the network and can be considered as the peripheral or clonal component. This is clearly very close to the architecture envisaged in the concept of second generation idiotypic networks (6–8) and similar to the idealized *ad hoc* architecture of (40) but in our model these properties evolve from simple principles.

In the steady state, the size of these groups and their linking does not change with time. The groups are built from clones of different idiotypes, which have an individual dynamics but share certain statistical properties. The building principles of these architectures can be described analytically (59, 60), and the statistical properties can be calculated within a mean-field theory in good agreement with simulations (61).

Whereas the preceding work by Brede and Behn (58), Schmidtchen and Behn (59), Schmidtchen et al. (60), and Schmidtchen and Behn (61) considered the autonomous idiotypic network, i.e., the network of B-lymphocytes and their antibodies without foreign or self antigen, we investigate here the evolution of the idiotypic network, in the presence of self, toward an architecture where the expansion of autoreactive clones is controlled by idiotypic interactions. Self is modeled by permanently present idiotypes which influence the evolution of the network but are themselves not affected by idiotypic interactions. Our model avoids the above reviewed drawbacks of previous attempts, and the results clearly support the view that the idiotypic network is instrumental in the control of autoreactive clones.

The paper is organized as follows. In Section 2, we describe essential features of the model, its update rules, the general building principles which allow to understand the structural properties of the expressed networks architecture, and a tool which allows a real time identification of patterns in simulations. In Section 3, we sketch the derivation of the mean-field theory which allows

to compute statistical properties if the structural properties of the pattern are known. In Section 4, we describe how the model should be modified in the presence of self. We report on simulations where the network in the presence of self evolves to an architecture such that the self is linked only to groups with very low population. Results of a modified mean-field theory are in good agreement with simulations. Finally, we give some conclusions and discuss problems for further research. There is a glossary where major key terms are briefly explained in a logical order.

2. THE MODEL

In this paper, we consider a minimal model of the idiotypic network (58), which is a coarse simplification of the real biological system but retains most important features and reveals a surprising complexity. The model has only few parameters and allows an analytical understanding of many of its properties.

2.1. POTENTIAL REPERTOIRE AND IDIOTYPIC INTERACTIONS

We model the repertoire of all possible idiotypes and their interactions by an undirected network, where each node ν of the network represents a distinct clone of B-lymphocytes of a given idiotype together with its antibodies. The idiotype is encoded by a bitstring of length d with entries 0 or 1. The number of different bitstrings 2^d is the size of the potential repertoire. Note that the bitstrings are not thought to represent the genetic code or the sequence of amino acids but are meant as a caricature of the phenotype allowing an easy notion of complementarity. Interpreting the entries of the bitstrings as coordinates in a d-dimensional space each node can be conceived as a corner of a d-dimensional unit hypercube.

B-lymphocytes receive a stimulus to proliferate if their receptors are cross-linked by complementary structures, which can be situated on antigen but also on antibodies of complementary idiotype. We represent possible idiotypic interactions by links between nodes of nearly complementary idiotype. Assuming only perfect complementary receptor structures seems unrealistic and it appears reasonable to allow small variations. Therefore, two nodes ν and u of our model are linked if their bitstrings are complementary allowing for up to m mismatches. We denote the undirected graph with 2^d nodes labeled by bitstrings of length d and links between complementary nodes with up to m mismatches as base graph $G_d^{(m)}$. Each node of the graph is linked to $\kappa = \sum_{k=0}^{m} \binom{d}{k}$ nodes, which we will call the neighborhood of a node in the following. For example, consider in $d = 12$ the bitstring **1 1 1 1 1 1 1 1 1 1 1 1**, which is perfect complementary to the bitstring **0 0 0 0 0 0 0 0 0 0 0 0**. Replacing anyone of the zero's in the latter by **1**, we obtain the 12 bitstrings which are complementary to the former except for one mismatch.

We only account whether an idiotypic clone is present or not and the corresponding node ν is either occupied $n(\nu) = 1$ or empty $n(\nu) = 0$. The subgraph of occupied nodes, the expressed repertoire, with its links represents the expressed idiotypic network at a certain time. In the following subsection, we describe how the expressed idiotypic repertoire is generated.

2.2. METADYNAMICS AND LOCAL DYNAMICS

There is a continuous influx of new B-lymphocytes from the bone marrow. There, by somatic random reshuffling of the VDJ genes, which are responsible for the binding sites of the variable regions of an antibody, different idiotypes of an enormous diversity are generated. The potential repertoire is estimated to exceed the order of 10^{10} (62). We model this metadynamics by occupying, in each step of an iteration procedure, empty nodes of the expressed network with probability p.

The stimulation of a B lymphocyte to proliferate is a non-monotonous, log-bell-shaped, function of the concentration of complementary structures (63). The number of cross-linked receptors increases with the concentration of complementary structures. However, if their concentration is too high, cross-linking becomes less likely due to a steric hindrance and the stimulation decreases. An unstimulated B-lymphocyte dies. In our model an occupied node, i.e., a clone of a certain idiotype only survives if the number of its occupied neighbors is in an allowed window between two thresholds, t_L and t_U. The survival of a clone depends in a deterministic way on its local neighborhood in the shape space.

The dynamics is described in discrete time, the time step should be chosen such that an unstimulated cell will die within this time span and a stimulated cell can proliferate. The temporal evolution of the network is induced by the following update rules:

(i) Influx: occupy empty nodes with probability p.
(ii) Window rule: count the number of occupied neighbors $n(\partial\nu)$ of node ν. If $n(\partial\nu)$ is outside the window $[t_L,t_U]$, set the node ν empty. This step is performed in parallel.
(iii) Iterate.

All three steps, the random global metadynamics, the deterministic local selection, and the iteration are of equal importance to describe an evolution of the network toward a complex architecture. Technically, our model can be categorized as a probabilistic cellular automaton, and also as a Boolean network, see Ref. (60) for a more detailed discussion.

Figure 1 illustrates the construction of the base graph and the application of the update rules for the case $d = 3$, $m = 1$, $[t_L, t_U] = [1,3]$.

Here, we report mainly on results for the following parameter setting, which is best investigated. The length of the bitstring is $d = 12$, then the network has $2^{12} = 4096$ nodes. We allow $m = 2$ mismatches, which make the linking neither too sparse nor too dense, each node has $\kappa = 79$ neighbors. The lower threshold t_L of the window rule has its minimal non-trivial value $t_L = 1$: for survival of a clone the stimulation by at least one anti-idiotypic clone is required. The upper threshold of the window rule is chosen as $t_U = 10$ that excludes very regular static patterns, which are in our context not interesting, for more details, see Ref. (60). Given these values, the influx probability p remains as main control parameter. In previous work (60, 61), we have studied a range for p from 0 to 0.1 and found that the architecture, which is of interest here evolves for p from 0.026 to 0.078. The results presented here explicitly are for p close to 0.078, where it is easier to initiate a reorganization of the pattern, but we have also studied a broader range of p. Simulations for longer bitstrings up to $d = 22$ have shown that many features are also found in larger networks and the major concepts of structural analysis are still applicable

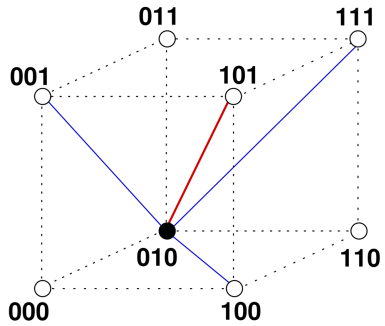


FIGURE 1 | Potential and expressed idiotypic repertoire. The figure shows all nodes whose idiotypes are encoded by bitstrings of length $d = 3$. The nodes can be thought as the corners of a 3-dimensional cube. Also shown are the links of node 010 (filled circle) to the node with perfect complementary idiotype 101 (red line), and to the nodes with idiotypes 100, 001, and 111, which are complementary allowing for one mismatch (blue lines). The dotted lines are only to visualize the edges of the cube. The base graph $G_d^{(m)}$ for $d = 3$ and $m = 1$ consists of these $2^d = 8$ nodes and all links connecting complementary nodes allowing for up to one mismatch. It represents the potential idiotypic repertoire and the possible idiotypic interactions. The influx of B cells from the bone marrow is modeled occupying empty nodes with probability p. Assume that node 010 is occupied. If the neighboring nodes representing clones of complementary idiotypes are empty, it receives no stimulation and will die. It will also die if there are too many, say all 4, nodes of complementary idiotype occupied. So the node 010 survives the update only if the number of occupied neighbor nodes is within a window $[t_L, t_U] = [1,3]$. These steps, random influx and parallel application of the window rule for all nodes, are repeated and lead for larger dimensions d and appropriate parameters to a steady state with a complex architecture of the expressed idiotypic network, see text.

(64). The program code is implemented in C++. For small base graphs ($d \approx 12$) optimization is not necessary. Larger base graphs ($d \approx 20$) require optimization and parallel computing.

2.3. BUILDING PRINCIPLES OF THE NETWORK ARCHITECTURE

Extensive simulations have shown that the network evolves, depending on the parameter choice, toward quasistationary states of possibly complex architecture (58). This architecture is characterized by groups of nodes that share statistical properties such as the mean occupation $\langle n(\nu)\rangle$ and the mean occupation of neighbors $\langle n(\partial\nu)\rangle$. The mean occupation of the nodes, the groups, and of the whole base graph, i.e., the size of the expressed idiotypic repertoire, all are stationary – which implies homeostasis. Although, the mean occupation of a single node is stationary, its actual occupation switches in time between 0 and 1. These switches are induced by both the random influx from the bone marrow and the deterministic window rule. A statistical characteristics of this behavior is the mean life time, which is also stationary and the same for all nodes of a group.

There are general building principles of the network's architecture which have been found by observing regularities in the bitstrings of nodes, which belong to the same group (59, 60). These principles make it possible to calculate the number of groups, their size, and the linkage between groups. Here, we only introduce the key terms and describe the essential results which are used in the following. For a deeper understanding of the derivation the reader should consult the original papers.

For a given architecture, the nodes can be classified according to their entries in the so-called determinant positions of the bitstrings. Different architectures have a different number $d_M \leq d$ of determinant positions. The group S_1 is defined as the set of all nodes with the same entries in all determinant positions, the entries in the non-determinant positions run through all 2^{d-d_M} possible combinations. Nodes in group S_2 differ in one determinant position compared to nodes in S_1, nodes in group S_3 in two determinant positions, and so on. Consequently we have $d_M + 1$ groups of size

$$|S_g| = 2^{d-d_M} \binom{d_M}{g-1} \tag{1}$$

for $g = 1, \ldots, d_M + 1$ and we can immediately observe that groups S_g and S_{d_M+2-g} have the same size.

The whole architecture can be build from smaller units, so-called pattern modules. These modules are the corners of a d_M-dimensional hypercube labeled by the determinant bits, together with the allowed links. Since the number of non-determinant bits is $d - d_M$, the whole architecture is obtained by arranging 2^{d-d_M} identical pattern modules and adding the allowed links between the nodes of these modules.

Next, we discuss the linkage of our idiotypic network in a pattern with d_M determinant bits on a base graph $G_d^{(m)}$. Each node in group S_i has a fixed number L_{ij} of links to nodes in group S_j. The L_{ij} are the elements of the link matrix $\mathbb{L}$. Since the update rule counts the number of occupied neighbors and all nodes of a group have the same mean occupation these data are of obvious interest to formulate a mean-field theory. A careful analysis of the bitstrings which encode the nodes of groups S_i and S_j allows to derive an explicit expression (59, 60) which can be written as

$$L_{ij} = \sum_{k=0}^{m} \sum_{r=0}^{k} \binom{i-1}{r} \binom{d_M - i + 1}{j - 1 - r} \times \binom{d - d_M}{k + j - 1 - 2r - (d_M - i + 1)}. \tag{2}$$

Given a pattern with d_M determinant bits there are $d_M + 1$ groups, therefore in equation (2) both i and j run from 1 to $d_M + 1$. As every node has κ neighbors, the row sum of $\mathbb{L}$ yields κ. Since $L_{ij} = L_{d_M+2-i,d_M+2-j}$ the link matrix is centrosymmetric, i.e., it fulfills the identity $\mathbb{L}J = J\mathbb{L}$ where the exchange matrix J has entries 1 on the counterdiagonal and 0 elsewhere. $\mathbb{L}$ describes a directed graph.

2.4. REAL TIME PATTERN IDENTIFICATION

In simulations, huge amounts of data are produced describing the occupation of each of the 2^d nodes of the network in every single time step. An enormous, namely logarithmic reduction of information can be reached by introducing a center of mass vector $\mathbf{R}$ in dimension d which allows a real time identification of patterns and detection of pattern changes (60). Instead of monitoring 2^d

data per time step it is enough to observe the d components of R. The center of mass vector is defined as

$$\mathbf{R} = \frac{1}{n(G)} \sum_{v} n(v)\mathbf{r}(v), \tag{3}$$

where the position vector $\mathbf{r}(v)$ of a node v, which is encoded by the bitstring $\mathbf{b}_d\mathbf{b}_{d-1}\cdots\mathbf{b}_1$ with $\mathbf{b}_i \in \{0,1\}$ has components $r_i(v) = 2\mathbf{b_i} - 1$. n(G) is the total occupation of the basegraph G. By definition, for a symmetrically occupied base graph, we have $\mathbf{R} = \mathbf{0}$, a symmetry breaking pattern is easy to identify.

In **Figure 2**, we see the time series of the components of **R** for the evolution toward a stationary 12-group pattern. The trajectory of R_2 fluctuates around zero. Since, the entries of non-determinant bits take for every group all possible values, and supposing that all nodes of a group are occupied with the same probability, the corresponding bit position can be identified as non-determinant. The trajectories of the five components R_7, R_9, R_{10}, R_{11}, R_{12} fluctuate around 0.4 and those of the 6 components R_1, R_3, R_4, R_5, R_6, R_8 around −0.4, The corresponding 11 bit positions are determinant. The dimension of the pattern module is $d_M = 11$, thus we have a 12-group architecture. Furthermore, we can readily identify the determinant bits of the group S_1. As explained below, S_1 is a peripheral group with high occupation. The observation that five components of **R** fluctuate around a positive value indicates that the five determinant bits at the corresponding positions should have entry 1, and the other six should have entry 0. Thus, the nodes of group S_1 have a bitstring $\mathbf{1\,1\,1\,1\,0\,1\,0\,0\,0\,0\cdot 0}$, where the $\cdot$ represents the only non-determinant bit. The determinant bits of S_{12} are complementary, and also nodes of the other groups are easily identified knowing their bitstrings. The reader who is interested in further technical details should consult (60).

The procedure is fast, robust against defects of patterns, and allows to identify pattern changes. Needless to say, the method hinges by construction on the encoding of the idiotype by bitstrings, which is only a gross caricature of the phenotype. Here, we use this tool to characterize the behavior of the network if several nodes become permanently occupied to mimic the presence of self.

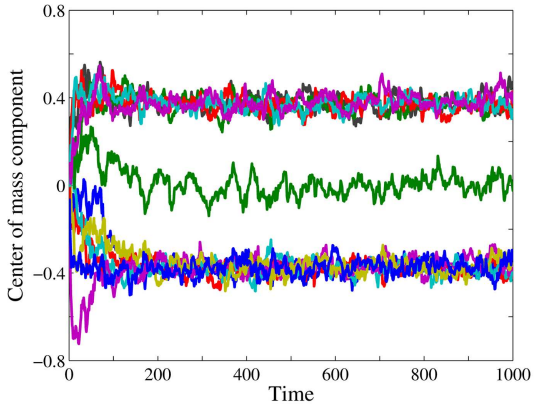


FIGURE 2 | Real time pattern identification. Time series of the components of the center of mass vector **R** given by equation (3), here on the base graph $G_{12}^{(2)}$ for a window $[t_L, t_U] = [1,10]$ and influx probability $p = 0.074$. Every color corresponds to one component of the center of mass vector. We start from an empty base graph, which is gradually occupied. A stationary state has evolved after about 200 time steps. The trajectory of R_2 (green) fluctuates around zero. The corresponding bit position is non-determinant, see text. The trajectories of the five components R_7, R_9, R_{10}, R_{11}, R_{12} fluctuate around 0.4 and those of the 6 components R_1, R_3, R_4, R_5, R_6, R_8 around −0.4, The corresponding bit positions are determinant. Together there are 11 determinant bits, hence the dimension of the pattern module is $d_M = 11$, and we can infer that the system has evolved toward a stationary 12-group architecture.

3. MEAN-FIELD THEORY

Once established, an architecture, characterized by the number of groups, their size, and their linking remains stationary for long periods of time and over some range of the main control parameter p. As shortly sketched above, for most architectures found in simulations their characteristics can be computed knowing the number of determinant bits d_M, which can be inferred from the time series of the center of mass coordinates.

The statistical properties of the nodes, which belong to the same group, such as the mean occupation and the mean life time, depend however on the actual value of p. They can be calculated (61) adopting the concept of mean-field theories, which was developed in statistical physics to describe phase transitions and has been transferred to many other problems in different fields. The main argument goes as follows.

The window rule (ii) for update of the occupation of a node counts only the total of the occupied neighbors. All nodes of a group have the same number of neighbors in the other groups given by the elements of the link matrix. The occupation of these neighbors typically fluctuates in time, and if they are many, it appears natural to replace the actual occupation of the neighbors by the average occupation. This works the better, the more neighbors are involved. In this view, a node feels only the different mean-fields, the modular mean-fields, exerted by the occupied neighboring nodes belonging to the different groups.

We now shortly describe the derivation in a more formal way to make the modifications understandable which are necessary when modeling the presence of self. Consider an architecture, which can be described by pattern modules of dimension d_M. Then, we have $d_M + 1$ groups of nodes S_g which share the mean occupation $\langle n(v_g)\rangle = n_g$ where $v_g \in S_g$, their linking is described by the link matrix $\mathbb{L}$. The set of mean occupations $\mathbf{n} = (n_1, \cdots, n_{d_M+1})^{\mathrm{T}}$ defines the state of the network in the reduced mean-field description at a certain time. Application of the update rules to **n** leads to a new state $\mathbf{n}'$ given by

$$\mathbf{n}' = \mathbf{f}(\mathbf{n}), \tag{4}$$

where the non-linear function **f** depends on the update rules and on the pattern we want to describe. We know that a node v_g of group S_g has L_{gl} neighbors in S_l. If the mean occupation in S_l is n_l, the new mean occupation after the influx with probability p is $\tilde{n}_l = n_l + p(1 - n_l)$. The probability that k_l nodes of the neighborhood in S_l are occupied after the influx is

$$\binom{L_{gl}}{k_l} \tilde{n}_l^{\,k_l}(1 - \tilde{n}_l)^{L_{gl}-k_l}. \tag{5}$$

Supposing that the groups are independent, the probability that for a micro-configuration with fixed $k_l, l = 1, \ldots, d_M + 1$, a total of $\sum_{l=1}^{d_M+1} k_l$ neighbors is occupied is simply the product of factors (5) for each group. Summing over all micro-configurations and taking into account the window rule leads to

$$\left[\sum_{k_l=0}^{L_{gl}}\right]_{l=1}^{d_M+1} \mathbb{1}(t_L \leq \sum_{l=1}^{d_M+1} k_l \leq t_U) \prod_{l=1}^{d_M+1} \binom{L_{gl}}{k_l} \tilde{n}_l^{k_l}(1-\tilde{n}_l)^{L_{gl}-k_l}, \tag{6}$$

where the indicator function $\mathbb{1}(\cdot)$ gives one, when the window rule in the parameters is fulfilled, otherwise zero. The last result should be multiplied with the mean occupation of a node of the considered group after the influx $\tilde{n}_g = n_g + p(1 - n_g)$ which gives

$$n'_g = \tilde{n}_g \left[\sum_{k_l=0}^{L_{gl}}\right]_{l=1}^{d_M+1} \mathbb{1}(t_L \leq \sum_{l=1}^{d_M+1} k_l \leq t_U) \times \prod_{l=1}^{d_M+1} \binom{L_{gl}}{k_l} \tilde{n}_l^{k_l}(1-\tilde{n}_l)^{L_{gl}-k_l}. \tag{7}$$

Iterating equation (7), for $g = 1, \ldots, d_M + 1$, the $\mathbf{n}'$ converge to a fixed point $\mathbf{n}^*$. Since $\mathbf{f}(\mathbf{n})$ is a non-linear function, several fixed points may exist. As a thumb rule, initial values close to the stationary average values seen in simulations are in the basin of attraction of fixed points of equation (7), which reproduce the simulation results. There may exist other fixed points, which were not found in simulations, for details see Ref. (61).

4. IDIOTYPIC NETWORK AND SELF

The 12-group architecture is of particular interest, as it strongly resembles the central and peripheral parts of the second generation idiotypic network. A scheme of these architecture is given in **Figure 3**. The 12-group architecture evolves on the base graph $G_{12}^{(2)}$ for $[t_L,t_U] = [1,10]$ and a range of p from 0.026 to 0.078. The groups comprise two self coupled core groups, two peripheral groups, which couple only to the core and five groups of stable holes. Stable holes are typically unoccupied since their occupied neighbors exceed t_U. Finally, there are three groups of singletons which are neighbored only by stable holes. Nodes of the singleton groups have an average occupation of 0.2–0.8, nodes of the periphery groups have 0.4–0.8 depending on p. The average occupation of the densely linked core groups is kept below 0.07, and the holes are almost empty, for details see Figure 8 in Ref. (61). Note that the singletons have no links to the connected part of the occupied network, which is built of the core and periphery groups. In terms of the second generation idiotypic networks, core, and periphery groups form the central part. The singletons, disconnected from the central part, form the clonal component (the peripheral part) of the second generation network.

The simplest possible way we can imagine to mimic the presence of self is to permanently occupy one or several nodes of the base graph and investigate their influence on the network architecture. The self nodes contribute to the number of occupied neighbors counted in the window rule but are themselves not affected by idiotypic interactions. The window rule does not apply to self nodes. We performed two types of computer experiments, inserting permanently occupied nodes in a fully developed 12-group architecture and monitoring the induced changes, or in an empty base graph and observing from scratch the evolution of the networks architecture.

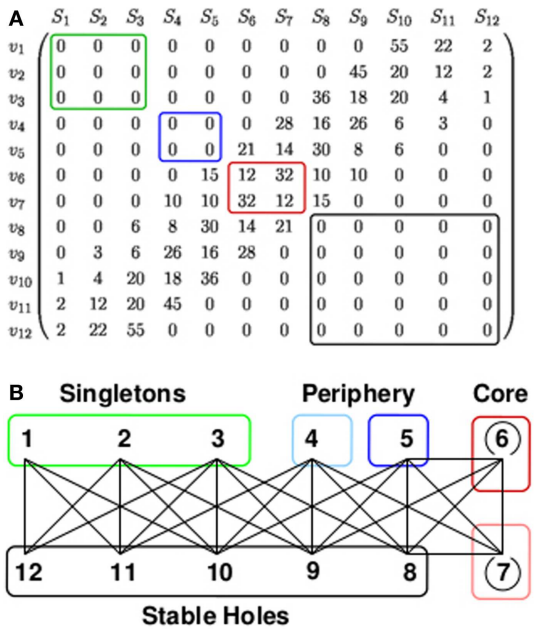


FIGURE 3 | 12-group architecture. **(A)** The entries L_{ij} of the link matrix, given by equation (2), show the number of neighbors a node v_i of group S_i has in group S_j. For example, the first row of the matrix tells that every node of the singleton group S_1 (denoted by v_1) is linked only to nodes of the hole groups S_{10}, S_{11}, and S_{12}, namely to 55, 22, and 2 nodes, respectively. Only nodes of S_6 and S_7 (red box) have links to other nodes of the own group. **(B)** The architecture generated by this link matrix together with a phenomenological classification into singletons (green), periphery (blue), core (red), and stable holes (black). The lines symbolize the existence of links between nodes of the connected groups, i.e., possible idiotypic interactions between the corresponding clones. The number of links which a node in S_i has to nodes in S_j is given by the element L_{ij} of the link matrix shown in **(A)**. The weakly occupied core groups have links within the own groups (open circles). The periphery groups are highly occupied and couple to the core and to the group of stable holes. The group of singletons is highly occupied and couples only to the stable holes. This architecture evolves on the base graph $G_{12}^{(2)}$ for a window $[t_L,t_U] = [1,10]$ and a range of the influx probability p from 0.026 to 0.078. See also Glossary.

4.1. SIMULATIONS

We performed extensive simulations for different protocols. Here, we describe only few most instructive cases.

We permanently occupy one node of the hole group S_{10} of an established 12-group pattern for $p = 0.076$, i.e., close but below the upper threshold of stability of the pattern. The hole groups have

many occupied neighbors and a self node staying there would be subject of a heavy autoimmune response. After few iterations, the former stable pattern destabilizes under the presence of the self and collapses. Thereafter, a new 12-group architecture evolves where the self node is now located in a group with only weakly occupied neighbors, which could be one of the singleton or periphery groups. **Figure 4A** shows the time series of the center of mass components for an example where the permanent occupied node (the self) is after a reorganization of the architecture finally in the periphery group S_5.

If we permanently occupy more than one node the scenario is similar. **Figure 4B** shows an example where we have permanently occupied 10 nodes of the hole group S_{10} of an established 12-group pattern for $p = 0.076$. The reorganization of the architecture is faster and in the new steady state the self nodes are found in singleton and periphery groups.

We also performed simulations where for a stationary 12-group pattern all members of the hole group S_{10} are permanently occupied. After reorganization of the architecture, in the steady state all self nodes belong to singletons and periphery groups and are never seen in a core or a hole group. If we start from an established 12-group pattern and permanently occupy one of the singletons or periphery groups this state will be stable for very long periods of time.

Starting from an empty base graph with several permanently occupied nodes, one observes that the architecture evolves from the very beginning such that the self nodes have only weakly occupied neighbors and thus are tolerated. This evolution from scratch toward a tolerant architecture occurs for a much broader range of p than the reorganization of an established architecture.

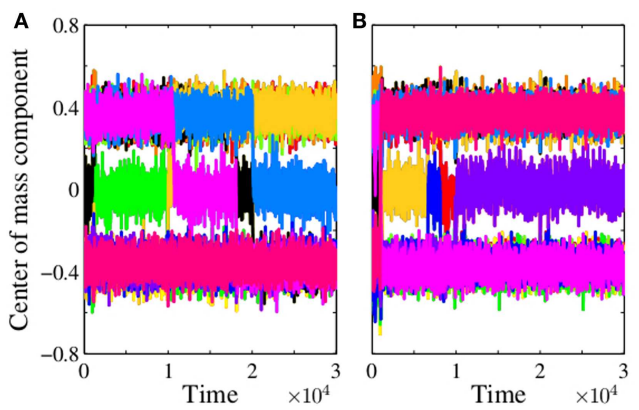


FIGURE 4 | Reorganization of the 12-group architecture with self. The figures displays the time series of the components of the center of mass vector obtained from simulations for an influx probability $p = 0.076$ when at $t = 0$ one node **(A)** or 10 nodes **(B)** of the hole group S_{10} are permanently occupied. Each of the 12 components is drawn with a different color. They are plotted one after another, only the last printed color is visible. The trajectories mainly fluctuate around ± 0.4 and zero. Jumps between these values indicate changes of the determinant bits associated with a reorganization of the architecture. For one self node **(A)** we see five jumps and for $t \gtrapprox 2 \times 10^4$ a stationary state is reached. For 10 self nodes **(B)**, after a few jumps, the stationary state is already reached for $t \gtrapprox 10^4$, obviously the impact of 10 self nodes is stronger than the impact of one. A closer look at the data (not discussed here) shows that the new stationary pattern has indeed a 12-group architecture where the self node in case **(A)** belongs to periphery group S_5 and in case **(B)** four self nodes belong to the singleton group S_3 and the remaining six self nodes belong to the periphery group S_5.

4.2. MEAN-FIELD THEORY WITH SELF

It is possible to modify the mean-field theory to describe a stationary architecture in the presence of self. We thus can describe situations where in an established pattern nodes are permanently occupied and the impact is so small that no reorganization sets in. If the impact is strong enough that a reorganization occurs and a new steady state emerges, we also can describe statistical properties of this steady state, such as the mean occupation of nodes and its neighbors, provided that we know its architecture.

We first consider one permanently occupied node of group S_s. It is linked to nodes of group S_g if $L_{sg} > 0$. The group S_g contains L_{sg} nodes that see the self. For these nodes we should modify the mean-field mapping, equation (7). The node of S_s which is permanently occupied should be exempted from the combinatorics of possible and allowed micro-configurations. Thus, we need to replace L_{gs} by $L_{gs} - 1$. Observe that $\binom{L_{gs}-1}{k_s}$ in the modified equation (7) is zero if $L_{gs} - 1$ is smaller than zero or k_s. To account for the permanently occupied self node, we should decrease both thresholds of the window condition by 1. For the $|S_g| - L_{sg}$ nodes of S_g which do not see the self node, the mapping is not modified. For example, for an influx with $p = 0.07$ and one permanently occupied node in a hole group or in a core group, $\langle n(\partial\nu)\rangle$ increases by about 1 and $\langle n(\nu)\rangle$ decreases by about 20% if ν is linked to the self node. The mean-field theory agrees with the simulation within 3–5%.

The case that all nodes of a group S_s are permanently occupied is even simpler because all nodes in group S_g see the same number L_{sg} of self nodes. We only have to modify the window condition decreasing both thresholds by L_{sg}. Note that if $t_U - L_{sg} < 0$ the modified window condition cannot be fulfilled and the indicator function $\mathbb{1}(\cdot)$ in the modified equation (7) returns 0. **Table 1** gives a detailed comparison of simulation and mean-field theory for the case that all 110 nodes of the singleton group S_{10}, cf. equation (1) for $d = 12$, $d_M = 11$, are occupied for $p = 0.074$.

For N_s self nodes with $1 < N_s < |S_s|$ the modification is also possible but more intricate and will not be reported here.

Encouraged by the good quantitative agreement between the steady states obtained in simulations and mean-field theory, we also looked at the time series of $\mathbf{n}$ generated by the mean-field mapping for a $d_M = 11$ pattern at $p = 0.074$ to see the effect induced by permanently occupying a group of nodes. We start with the fixed point $\mathbf{n}^\star$, which describes a 12-group pattern where the groups are ordered as in **Figure 5A**. In the steady state, at an arbitrary iteration step, we permanently occupy the hole group S_{10}. The time series, cf. **Figure 6**, shows that this state immediately destabilizes and that a reorganization sets in. The pattern converges to a new state where the self belongs to the new singleton group S_{10}. These singletons have only neighbors in the new unoccupied hole groups, see **Figure 5B**. The network controls the expansion of the autoreactive idiotypes in the hole groups – thus providing self-tolerance. Analogous results (not shown here) are obtained if we permanently occupy the hole group S_9, after reorganization

Table 1 | 12-Group architecture with self after reorganization.

Group	$\langle n(v)\rangle$		$\langle n(\partial v)\rangle$	
	Simulation	MFT	Simulation	MFT
S_1	0.0	0.0	71.75 (54.01)	71.40 (53.99)
S_2	0.0	0.0	60.34 (53.86)	60.29 (53.96)
S_3	0.0	0.0	59.62 (53.50)	59.85 (53.52)
S_4	0.0	0.0	36.70 (34.86)	36.62 (34.72)
S_5	0.0	0.0	31.5 (29.62)	31.63 (29.73)
S_6	0.002 (0.001)	0.0	13.52 (13.53)	13.63 (13.63)
S_7	0.01	0.003	10.12 (10.09)	10.10
S_8	0.677 (0.661)	0.6708	0.15 (0.14)	0.07
S_9	0.706 (0.681)	0.6827	0.025 (0.018)	0.01
S_{10}	1.0 (0.685)	1.0 (0.685)	0.02 (0.0)	0.0
S_{11}	0.685 (0.684)	0.6835 (0.685)	0.001 (0.0)	0.0
S_{12}	0.685 (0.682)	0.6835 (0.685)	0.001 (0.0)	0.0

*The 110 nodes of the singleton group S_{10} are permanently occupied to mimic the presence of self antigen, see **Figure 5B**. The table shows the mean occupation $\langle n(v)\rangle$ and the mean occupation of neighbors $\langle n(\partial v)\rangle$ for all groups as obtained for $p = 0.074$ from simulations and from mean-field theory (MFT) with a $d_M = 11$ module. When deviating, the data for the case without self are given in parentheses. The groups $S_1, \ldots, S_5$ have direct neighbors in S_{10}, where S_1 has the most ones. Therefore, the change in $\langle n(\partial v)\rangle$ due to self is largest for S_1. Results from simulation and mean-field theory are in good agreement. The simulation data are obtained as follows. We first computed the temporal average of each node's occupation from 30,000 time steps. Then the mean of these data over all nodes of the same group is calculated. The variance of the mean over the group members is of the order 10^{-3}.*

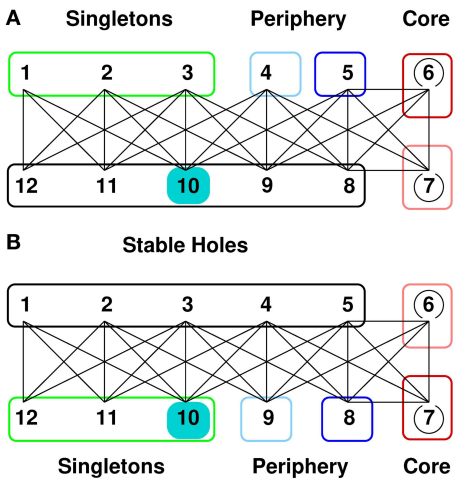


FIGURE 5 | 12-group architecture with self. **(A)** We permanently occupy one of the hole groups, group 10 (cyan), thus mimicking the permanent presence of self. This state is not favorable since the self couples to singletons and periphery, which have a high occupation. **(B)** Letting the thus prepared system evolve, it soon reaches a new steady state, still a 12-group architecture, but organized such that the self now belongs to the singletons and thus couples only to the almost empty stable holes. The self-recognizing idiotypes are controlled by the network, thus providing self-tolerance.

group S_9 is a periphery group coupling only to the holes and to the weakly occupied core.

We note in this context that due to the centrosymmetry of the link matrix of the autonomous network without self, given a fixed point $\mathbf{n}^* = (n_1^*, n_2^*, \ldots, n_{d_M+1}^*)^{\mathrm{T}}$, there exists always a mirrored fixed point $\mathbf{n}^*_{\mathrm{mirror}} = (n_{d_M+1}^*, \ldots, n_2^*, n_1^*)^{\mathrm{T}}$. Obviously this symmetry is broken if self is present.

5. CONCLUSION AND OUTLOOK

We have extended a minimal model of the idiotypic network (58, 60, 61) to study the evolution of the network in the presence of self. Self is represented by permanently occupied nodes of certain idiotypes. These self nodes can stimulate autoreactive clones and thus influence the evolution of the network but are themselves not affected by the idiotypic interactions. We report on simulation results for the case that the self nodes are permanently occupied already at the initial state. Then, the network evolves toward an architecture where the permanently occupied self nodes are incorporated into groups of nodes which have, in a sense, a similar idiotype. These groups can idiotypically interact only with other groups that are either completely suppressed by the network (stable holes) or only weakly occupied. The network controls the expansion of self-reactive clones thus providing self-tolerance.

We also studied the response of a network with an already established architecture to a sudden appearance of self antigen. Nodes of the hole groups were permanently occupied, which is most unfavorable since these groups are linked to highly occupied clones. Provided that the influx from the bone marrow is sufficiently high the network reorganizes its architecture such that in the end the self nodes belong to groups, which have only empty ore weakly occupied neighbors, as in the previous case.

For the simplest cases that only one node or all nodes of a group are permanently occupied, we have modified the mean-field theory and found good agreement of analytical and simulation results.

As discussed in the introduction to some extent, there are preceding attempts in the literature, which aim in the same direction but were not really satisfying. Our results strongly support the view that idiotypic interactions can be instrumental in the control of autoreactive clones.

The network in the presence of self has been previously studied by one of us in simulations for one self node on the base graph $G_{12}^{(3)}$ with weighted links. The weights were given according to the number of mismatches of the linked nodes and the window condition was modified accordingly. The patterns are slightly easier to destabilize which explains why the phenomenon of self-tolerance was first observed in that version of the model (65).

Further studies should systematically explore the system's behavior for other protocols, e.g., for arbitrary numbers of self nodes possibly distributed over the whole base graph, desirably in both simulations and an accordingly extended mean-field approach.

It is of obvious interest to investigate in the frame of the model possible reasons for failure of self-tolerance. Transitions from a healthy self-tolerant state to an autoimmune state by a perturbation, possibly an ordinary infection, of the clones that

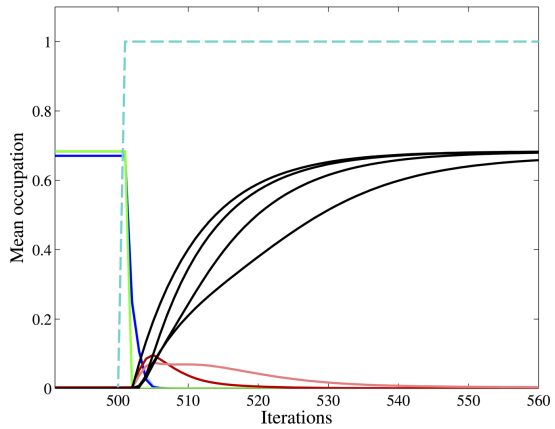


FIGURE 6 | Mean-field theory of a 12-group architecture with self. The figure shows the time series of the mean occupation per node of the 12 groups as obtained from iterating equation (7) for an influx probability $p = 0.074$. We start with the autonomous system in the steady state where singletons (green) and periphery (blue) have a mean occupation per node $\langle n \rangle \approx 0.68$, whereas core (red) and stable holes (black) have $\langle n \rangle \approx 0$. The color code is the same as in **Figure 5A**. At some arbitrary iteration step (here 500), we permanently occupy the 110 nodes of the hole group S_{10} (dashed cyan line). The fixed point of equation (7) looses its stability and a new mirrored architecture emerges where the permanently occupied nodes, the self, now belong to the singletons, which have neighbors only in the empty hole groups, cf. **Figure 5B**. The occupation of the previous singleton (green line) and periphery (blue line) groups drops down to almost zero, whereas the previous hole groups (black lines) become occupied as typical for singletons and periphery. (The light blue line of the periphery group S_4 is not visible here since it is covered by the green line up to iteration step 500, and thereafter by the blue line.) After an temporary increase the core groups (red lines) return to its previous occupation. For a detailed comparison of steady state results of mean-field theory and simulation see **Table 1**.

control the autoreactive idiotypes should be considered, together with the reverse phenomenon of 'spontaneous' remission from an autoimmune to a healthy state. Therapeutic strategies adopting the network paradigm (66), which consist in stimulating the protective clones that control the autoreactive clones, instead of applying immunosuppressive drugs, could be modeled.

To study age induced effects, it would be very interesting to consider an influx rate p from the bone marrow, which decreases over the lifespan of an individual. The architecture which controls autoreactive clones is found for a certain range of p. However, if we suddenly stop the influx at all, the group of singletons, which is only sustained by the influx would be depopulated, and the other, connected part of the network would, in a sense, freeze. Since the singletons play an important part in controlling the autoreactive clones this should have consequences for maintaining self-tolerance. A small influx outside the range mentioned above would lead to less complex architectures which may be not functional.

The renewal rate of the expressed idiotypic repertoire is certainly relevant in the physiological context. Therefore, it would be interesting to determine this characteristics in the frame of our model. It is of course related to the influx from the bone marrow but also depending on the population dynamics of the B-cell clones. It would be collective characteristics of a group and is more difficult to determine than the mean life of a single clone.

Motivation to develop our mathematical model further comes also from experimental and clinical medicine and from the progress of microarray technologies.

Hampe (10) reviewed the role of anti-idiotypic antibodies in autoimmunity, including Type 1 Diabetes. There is experimental evidence of anti-Id mediated neutralization of autoantibodies, e.g., in Myasthenia gravis, or suppression of autoantibody secretion, e.g., in Idiopathic thrombocytopenic purpura. For a number of autoimmune diseases including systemic lupus erythematosus and autoimmune thyroid diseases it has been shown that anti-Id specific to autoantibodies are present in patients during remission and/or in healthy individuals, whereas it is absent during periods of active disease. The formation of anti-Id-autoantibody complexes makes it difficult to detect the single constituents by conventional assays, but several methods have been developed to overcome this problem.

Monoclonal antibodies become rapidly important in clinical therapies of autoimmune and inflammatory diseases, see e.g., Ref. (67). This gives a strong motivation to improve our understanding of systemic consequences of immunomanipulation.

The rapid technological progress makes large scale studies of the expressed idiotypic repertoire feasible. The use of antigen microarrays to profile the autoantibody repertoire in health and disease is reviewed in Ref. (68), for an application of network theory to detect antibody trees associated with antigen see (69). Immunosignaturing, reviewed in Ref. (70) uses random-sequence peptide microarrays. Microarrays using antibodies or proteins are however still expensive and complicated (70). In addition, inferring the network architecture from a sample, which is only a snapshot of a subset of the expressed idiotypic repertoire is a very demanding task.

Our model, which provides an analytical understanding of the network architecture, could be helpful to formulate conditions for a new generation of experiments with the aim to infer the networks architecture and to elucidate its role in healthy conditions and disease. From the viewpoint of statistical physics or systems biology, the question appears natural and most interesting whether there is a general principle which guides the evolution of the idiotypic network.

6. GLOSSARY

- **Nodes:** A node of the network represents a clone of B-lymphocytes of a certain idiotype together with its antibodies. At a given time a node can be either occupied or empty, the corresponding clone is present or absent, respectively.
- **Bitstrings:** An idiotype is encoded by a bitstring of length d with entries 0 or 1. There is a total of 2^d different bitstrings, which is the size of the potential idiotypic repertoire.
- **Links:** A link of the network connects two nodes with complementary idiotype, i.e., with complementary bitstrings. We do not require perfect complementarity but allow for up to m mismatches. The links represent the possible idiotypic interactions between the clones of the potential idiotypic repertoire.
- **Base graph:** The base graph consists of all nodes and their links for a given choice of d and m. It represents the potential idiotypic

repertoire and the possible idiotypic interactions, which take place if the linked nodes are occupied.

- **Influx:** The influx of new B-lymphocytes with random idiotype from the bone marrow is modeled by occupying empty nodes with a certain probability p in each iteration of an update procedure.
- **Window rule:** The window rule decides whether an occupied node will survive the update or not. It will only remain occupied if the number of its occupied neighbors is neither too small nor too high but lies within an allowed window with lower and upper thresholds, t_L and t_U. The window rule is applied in parallel for all nodes of the network in each iteration of an update.
- **Evolution:** Iterating the steps of random innovation (influx) and deterministic selection (window rule) induces an evolution which leads, after a transient period, toward a quasistationary state of the network which may have, depending on the parameter setting, a very complex architecture.
- **Architecture:** In the steady state, groups of nodes can be identified which share statistical properties such as mean occupation and mean occupation of neighbors. The number of groups, their size, and their linking remain constant and characterize the architecture. The most interesting architecture, considered in this paper, comprises a connected part (core and periphery), a hereof disconnected part (singletons), and groups of suppressed clones (stable holes).
- **Pattern modules:** The architecture can be build by arranging identical smaller units, the pattern modules which are constructed like the base graph but have a smaller dimension d_M. A pattern module contains at least one node from every group. Given d_M, the number of groups, their size, and their linking can be calculated.
- **Stable Holes:** Group of nodes that are mainly unoccupied because the number of their occupied neighbors typically exceeds by far the upper threshold of the window rule, therefore we call this group stable holes. The mean occupation is close to zero.
- **Core:** Groups consisting of nodes with links to nodes in the same group build the core. The mean occupation is very low.
- **Periphery:** Groups consisting of nodes linked to the core and to stable holes, but not to nodes in its own group. The core and periphery correspond to the central part of the network. The mean occupation is high.
- **Singletons:** Groups of nodes that are only connected to stable holes. An occupied singleton can survive if it has, after the influx step, an occupied neighbor (in the group of stable holes) which typically does not survive applying the window rule. The mean occupation is high.
- **Mean-field theory:** The mean-field theory allows for a given architecture to calculate statistical properties of the groups, independent of simulations. The main simplification is that the actual occupation of neighboring nodes is replaced by their mean occupation which works the better the more neighbors are involved.
- **Self:** In the extended version of the model, self is represented by permanently occupied nodes of the network that exert influence on the linked neighbor nodes but are themselves not affected by idiotypic interactions. The window rule does not apply to self nodes.

ACKNOWLEDGMENTS

The authors thank Holger Schmidtchen and Rüdiger Kürsten for valuable discussions, and Nicolas Preuß for carefully checking the numerical results in **Table 1**.

REFERENCES

1. Coutinho A. Beyond clonal selection and network. *Immunol Rev* (1989) **110**(1):63–88. doi:10.1111/j.1600-065X.1989.tb00027.x
2. Jerne NK. Towards a network theory of the immune system. *Ann Immunol* (1974) **125C**:373–89.
3. Jerne NK. Idiotypic networks and other preconceived ideas. *Immunol Rev* (1984) **79**(1):5–24. doi:10.1111/j.1600-065X.1984.tb00484.x
4. Jerne NK. The generative grammar of the immune system. *EMBO J* (1985) **4**:847–52.
5. Sim GK, MacNeil IA, Augustin A. T helper cell receptors: idiotypes and repertoire. *Immunol Rev* (1986) **90**:49–72. doi:10.1111/j.1600-065X.1986.tb01477.x
6. Huetz F, Jacquemart F, Peña Rossi C, Varela F, Coutinho A. Autoimmunity: the moving boundaries between physiology and pathology. *J Autoimmun* (1988) **1**(6):507–18. doi:10.1016/0896-8411(88)90044-3
7. Varela FJ, Coutinho A. Second generation immune networks. *Immunol Today* (1991) **12**(5):159–66. doi:10.1016/S0167-5699(05)80046-5
8. Coutinho A. A walk with Francisco Varela from first- to second-generation networks: in search of the structure, dynamics and metadynamics of an organism-centered immune system. *Biol Res* (2003) **36**(1):17–26. doi:10.4067/S0716-97602003000100004
9. Urbain J, Wikler M, Franssen J, Collignon C. Idiotypic regulation of the immune system by the induction of antibodies against anti-idiotypic antibodies. *Proc Natl Acad Sci U S A* (1977) **74**(11):5126–30. doi:10.1073/pnas.74.11.5126
10. Hampe CS. Protective role of anti-idiotypic antibodies in autoimmunity – lessons for type 1 diabetes. *Autoimmunity* (2012) **45**(4):320–31. doi:10.3109/08916934.2012.659299
11. Avrameas S. Natural autoantibodies: from 'horror autotoxicus' to 'gnothi seauton'. *Immunol Today* (1991) **12**(5):154–9. doi:10.1016/0167-5699(91)90080-D
12. Shoenfeld Y, George J. Induction of autoimmunity. A role for the idiotypic network. *Ann N Y Acad Sci* (1997) **815**:342–9. doi:10.1111/j.1749-6632.1997.tb52080.x
13. Shoenfeld Y. The idiotypic network in autoimmunity: antibodies that bind antibodies that bind antibodies. *Nat Med* (2004) **10**(1):17–8. doi:10.1038/nm0104-17
14. Pendergraft WF, Preston GA, Shah RR, Tropsha A, Carter CW, Jennette JC, et al. Autoimmunity is triggered by cPR-3(105-201), a protein complementary to human autoantigen proteinase-3. *Nat Med* (2003) **10**(1):72–9. doi:10.1038/nm968
15. McGuire KL, Holmes DS. Role of complementary proteins in autoimmunity: an old idea re-emerges with new twists. *Trends Immunol* (2005) **26**(7):367–72. doi:10.1016/j.it.2005.05.001
16. Tzioufas AG, Routsias JG. Idiotype, anti-idiotype network of autoantibodies. *Autoimmun Rev* (2010) **9**(9):631–3. doi:10.1016/j.autrev.2010.05.013
17. Routsias JG, Tzioufas AG. B-cell epitopes of the intracellular autoantigens Ro/SSA and La/SSB: tools to study the regulation of the autoimmune response. *J Autoimmun* (2010) **35**(3):256–64. doi:10.1016/j.jaut.2010.06.016
18. Dwyer DS, Vakil M, Kearney J. Idiotypic network connectivity and a possible cause of myasthenia gravis. *J Exp Med* (1986) **164**(4):1310–8. doi:10.1084/jem.164.4.1310
19. Langman R, Cohn M. Editorial introduction. *Semin Immunol* (2000) **12**(3):159–62. doi:10.1006/smim.2000.0227
20. Langman R, Cohn M. A minimal model for the self-nonself discrimination: a return to the basics. *Semin Immunol* (2000) **12**(3):189–95. doi:10.1006/smim.2000.0231
21. Anderson CC, Matzinger P. Danger: the view from the bottom of the cliff. *Semin Immunol* (2000) **12**(3):231–8. doi:10.1006/smim.2000.0236
22. Grossman Z, Paul WE. Self-tolerance: context dependent tuning of T cell antigen recognition. *Semin Immunol* (2000) **12**(3):197–203. doi:10.1006/smim.2000.0232
23. Grossman Z, Paul WE. Autoreactivity, dynamic tuning and selectivity. *Curr Opin Immunol* (2001) **13**(6):687–98. doi:10.1016/S0952-7915(01)00280-1
24. Grossman Z, Min B, Meier-Schellersheim M, Paul WE. Opinion: concomitant regulation of T-cell activation and homeostasis. *Nat Rev Immunol* (2004) **4**(5):387–95. doi:10.1038/nri1355

25. Cohen I. Discrimination and dialogue in the immune system. *Semin Immunol* (2000) **12**(3):215–9. doi:10.1006/smim.2000.0234
26. Cohen IR. Biomarkers, self-antigens and the immunological homunculus. *J Autoimmun* (2007) **29**(4):246–9. doi:10.1016/j.jaut.2007.07.016
27. Poletaev AB, Stepanyuk VL, Gershwin EM. Integrating immunity: the immunculus and self-reactivity. *J Autoimmun* (2008) **30**(1-2):68–73. doi:10.1016/j.jaut.2007.11.012
28. Zinkernagel R. Localization dose and time of antigens determine immune reactivity. *Semin Immunol* (2000) **12**(3):163–71. doi:10.1006/smim.2000.0253
29. Sakaguchi S, Wing K, Miyara M. Regulatory T cells – a brief history and perspective. *Eur J Immunol* (2007) **37**(S1):S116–23. doi:10.1002/eji.200737593
30. Sakaguchi S, Wing K, Onishi Y, Prieto-Martin P, Yamaguchi T. Regulatory T cells: how do they suppress immune responses? *Int Immunol* (2009) **21**(10):1105–11. doi:10.1093/intimm/dxp095
31. Shoenfeld Y, Rose NR editors. *Infection and Autoimmunity*. Amsterdam: Elsevier (2004).
32. Vanderlugt CL, Miller SD. Epitope spreading in immune-mediated diseases: implications for immunotherapy. *Nat Rev Immunol* (2002) **2**(2):85–95. doi:10.1038/nri724
33. Meda F, Folci M, Baccarelli A, Selmi C. The epigenetics of autoimmunity. *Cell Mol Immunol* (2011) **8**(3):226–36. doi:10.1038/cmi.2010.78
34. Zouali M editor. *The Epigenetics of Autoimmune Diseases*. Chichester: Wiley (2009).
35. Wucherpfennig KW, Noel RJ. Editorial overview. *Curr Opin Immunol* (2011) **23**(6):699–701. doi:10.1016/j.coi.2011.10.003
36. Pillai S, Mattoo H, Cariappa A. B cells and autoimmunity. *Curr Opin Immunol* (2011) **23**(6):721–31. doi:10.1016/j.coi.2011.10.007
37. Fagarasan S, Honjo T. T-independent immune response: new aspects of B cell biology. *Science* (2000) **290**(5489):89–92. doi:10.1126/science.290.5489.89
38. Bogen B, Ruffini P. Review: to what extent are T cells tolerant to immunoglobulin variable regions? *Scand J Immunol* (2009) **70**(6):526–30. doi:10.1111/j.1365-3083.2009.02340.x
39. Salinas GF, Braza F, Brouard S, Tak P-P, Baeten D. The role of B lymphocytes in the progression from autoimmunity to autoimmune disease. *Clin Immunol* (2013) **146**(1):34–45. doi:10.1016/j.clim.2012.10.005
40. Stewart J, Varela FJ. Exploring the meaning of connectivity in the immune network. *Immunol Rev* (1989) **110**:37–61. doi:10.1111/j.1600-065X.1989.tb00026.x
41. Kearney J, Vakil M, Nicholson A. Non-random VH gene expression and idiotype anti-idiotype expression in early B cells. In: Kelsoe G, Schulze D editors. *Evolution and Vertebrate Immunity: The Antigen Receptor and MHC Gene Families*. Austin: Texas University Press (1987). p. 175–90.
42. Stewart J, Varela FJ, Coutinho A. The relationship between connectivity and tolerance as revealed by computer simulation of the immune network: some lessons for an understanding of autoimmunity. *J Autoimmun* (1989) **2**:15–23. doi:10.1016/0896-8411(89)90113-3
43. Sulzer B, van Hemmen JL, Behn U. Central immune system, the self and autoimmunity. *Bull Math Biol* (1994) **56**(6):1009–40. doi:10.1016/S0092-8240(05)80331-3
44. Calenbuhr V, Varela F, Bersini H. Natural tolerance as a function of network connectivity. *Int J Bifurcat Chaos* (1996) **06**(09):1691–702. doi:10.1142/S0218127496001041
45. De Boer RJ, Perelson AS. Size and connectivity as emergent properties of a developing immune network. *J Theor Biol* (1991) **149**(3):381–424. doi:10.1016/S0022-5193(05)80313-3
46. Stewart J, Varela FJ. Morphogenesis in shape-space. Elementary meta-dynamics in a model of the immune network. *J Theor Biol* (1991) **153**(4):477–98. doi:10.1016/S0022-5193(05)80152-3
47. Takumi K, De Boer RJ. Self assertion modeled as a network repertoire of multideterminant antibodies. *J Theor Biol* (1996) **183**:55–66. doi:10.1006/jtbi.1996.0201
48. Carneiro J, Coutinho A, Faro J, Stewart J. A model of the immune network with B-T cell co-operation. I – prototypical structures and dynamics. *J Theor Biol* (1996) **182**(4):513–29. doi:10.1006/jtbi.1996.0193
49. Carneiro J, Coutinho A, Stewart J. A model of the immune network with B-T cell co-operation. II – the simulation of ontogenesis. *J Theor Biol* (1996) **182**(4):531–47. doi:10.1006/jtbi.1996.0193
50. Leon K, Carneiro J, Perez R, Montero E, Lage A. Natural and induced tolerance in an immune network model. *J Theor Biol* (1998) **193**(3):519–34. doi:10.1006/jtbi.1998.0720
51. Stewart J, Coutinho A. The affirmation of self: a new perspective on the immune system. *Artif Life* (2004) **10**:261–76. doi:10.1162/1064546041255593
52. Behn U. Idiotypic networks: toward a renaissance? *Immunol Rev* (2007) **216**(1):142–52. doi:10.1111/j.1600-065X.2006.00496.x
53. Behn U. Idiotype network. In: Delves P, editor. *Encyclopedia of Life Sciences (ELS)*. Chichester: John Wiley & Sons, Ltd (2011). p. 1–11. doi:10.1002/9780470015902.a0000954.pub2
54. Bersini H. Self-assertion versus self-recognition: a tribute to Francisco Varela. In: Timmis J, Bentley PJ editors. *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*. Canterbury: University of Kent at Canterbury Printing Unit (2002). p. 107–112.
55. Vaz NM. Francisco Varela and the immunological self. *Syst Res Behav Sci* (2011) **28**(6):696–703. doi:10.1002/sres.1126
56. Vaz NM. The specificity of immunologic observations. *Construct Found* (2011) **6**(3):334–42. Available from: http://www.univie.ac.at/constructivism/journal/6/3/334.vaz
57. Vaz NM, Ramos GC, deCastro AB. The enactive paradigm 33 years later. Response to Alfred Tauber. *Construct Found* (2011) **6**(3):345–51. Available from: http://www.univie.ac.at/constructivism/journal/6/3/345.vaz
58. Brede M, Behn U. Patterns in randomly evolving networks: idiotypic networks. *Phys Rev E* (2003) **67**(3):031920. doi:10.1103/PhysRevE.67.031920
59. Schmidtchen H, Behn U. Randomly evolving idiotypic networks: analysis of building principles. In: Bersini H, Carneiro J editors. *Artificial Immune Systems, Volume 4163 of Lecture Notes in Computer Science*. Berlin: Springer (2006). p. 81–94.
60. Schmidtchen H, Thüne M, Behn U. Randomly evolving idiotypic networks: structural properties and architecture. *Phys Rev E* (2012) **86**(1):011930. doi:10.1103/PhysRevE.86.011930
61. Schmidtchen H, Behn U. Randomly evolving idiotypic networks: modular mean field theory. *Phys Rev E* (2012) **86**(1):011931. doi:10.1103/PhysRevE.86.011931
62. Berek C, Milstein C. The dynamic nature of the antibody repertoire. *Immunol Rev* (1988) **105**:5–26. doi:10.1111/j.1600-065X.1988.tb00763.x
63. Perelson A, Weisbuch G. Immunology for physicists. *Rev Mod Phys* (1997) **69**:1219. doi:10.1103/RevModPhys.69.1219
64. Sachsenweger H. *Bitstring Model for the Idiotypic Network: Automatized Pattern Identification and Simulation of Large Systems*. Master's thesis. Leipzig: University Leipzig, Institute for Theoretical Physics (2012).
65. Werner B. *Idiotypische Netzwerke mit Antigenen: Gedächtnis und Selbsttoleranz*. Diplomarbeit. Leipzig: University Leipzig, Institute for Theoretical Physics (2010).
66. Feldmann M, Steinman L. Design of effective immunotherapy for human autoimmunity. *Nature* (2005) **435**(7042):612–9. doi:10.1038/nature03727
67. Chan AC, Carter PJ. Therapeutic antibodies for autoimmunity and inflammation. *Nat Rev Immunol* (2010) **10**:301–16. doi:10.1038/nri2761
68. Cohen IR. Autoantibody repertoires, natural biomarkers, and system controllers. *Trends Immunol* (2013) **34**(12):620–5. doi:10.1016/j.it.2013.05.003
69. Madi A, Kenett DY, Bransburg-Zabary S, Merbl Y, Quintana FJ, Tauber AI, et al. Network theory analysis of antibody-antigen reactivity data: the immune trees at birth and adulthood. *PLoS One* (2011) **6**(3):e17445. doi:10.1371/journal.pone.0017445
70. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical characterization of the "immunosignaturing effect." *Mol Cell Proteom* (2012) **11**(4):M111.011593. doi:10.1074/mcp.M111.011593

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; accepted: 19 February 2014; published online: 10 March 2014.
*Citation: Schulz R, Werner B and Behn U (2014) Self-tolerance in a minimal model of the idiotypic network. Front. Immunol. **5**:86. doi: 10.3389/fimmu.2014.00086*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Schulz, Werner and Behn. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 03 June 2014
doi: 10.3389/fimmu.2014.00264

Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas

Miri Michaeli[1†], Hilla Tabibian-Keissar[1,2†], Ginette Schiby[2†], Gitit Shahaf[1], Yishai Pickman[1], Lena Hazanov[1], Kinneret Rosenblatt[2], Deborah K. Dunn-Walters[3], Iris Barshack[2,4‡] and Ramit Mehr[1*‡]

[1] *The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel*
[2] *Department of Pathology, Sheba Medical Center, Ramat Gan, Israel*
[3] *Division of Immunology, Infection, and Inflammatory Diseases, King's College London School of Medicine, London, UK*
[4] *Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Jose Faro, Universidade de Vigo, Spain
Andrew M. Collins, University of New South Wales, Australia

****Correspondence:***
Ramit Mehr, The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel
e-mail: ramit.mehr@biu.ac.il

[†] Miri Michaeli, Hilla Tabibian-Keissar and Ginette Schiby have contributed equally to this work.

[‡] Iris Barshack and Ramit Mehr have equally supervised this study.

Chronic gastritis is characterized by gastric mucosal inflammation due to autoimmune responses or infection, frequently with *Helicobacter pylori*. Gastritis with *H. pylori* background can cause gastric mucosa-associated lymphoid tissue lymphoma (MALT-L), which sometimes further transforms into diffuse large B-cell lymphoma (DLBCL). However, gastric DLBCL can also be initiated *de novo*. The mechanisms underlying transformation into DLBCL are not completely understood. We analyzed immunoglobulin repertoires and clonal trees to investigate whether and how immunoglobulin gene repertoires, clonal diversification, and selection in gastritis, gastric MALT-L, and DLBCL differ from each other and from normal responses. The two gastritis types (positive or negative for *H. pylori*) had similarly diverse repertoires. MALT-L dominant clones (defined as the largest clones in each sample) presented higher diversification and longer mutational histories compared with all other conditions. DLBCL dominant clones displayed lower clonal diversification, suggesting the transforming events are triggered by similar responses in different patients. These results are surprising, as we expected to find similarities between the dominant clones of gastritis and MALT-L and between those of MALT-L and DLBCL.

Keywords: B-cells, gastritis, *H. pylori*, MALT lymphoma, DLBCL, Ig gene, repertoire, somatic hypermutation

INTRODUCTION

Chronic gastritis is a common disorder characterized by chronic inflammation of gastric mucosa. In acute gastritis, patients suffer from dyspeptic symptoms including epigastric burning, distention or bloating, belching, episodic nausea, flatulence, and halitosis. In contrast, most patients with chronic gastritis are asymptomatic (1). One of the major causes of gastritis is bacterial infection, most frequently with *Helicobacter pylori* (*H. pylori*). *H. pylori* are Gram-negative bacteria that are present in the gastric mucosa of more than 50% of people and may persist lifelong unless treated (2). *H. pylori* are resistant to elimination by the immune response so the immune system fails to remove the infection effectively (3). Previous studies have shown a strong association between gastritis and *H. pylori* infection, at least in the early stages of gastritis (3, 4). Although rare, organisms other than *H. pylori* (e.g., *Mycobacterium avium*-intracellulare, Herpes simplex, Cytomegalovirus, and Epstein–Barr virus) can invade the gastric mucosa and cause inflammation (5, 6). Gastritis can also be initiated *de novo*, as an autoimmune disease (7). In either case, prolonged antigenic stimulation causing chronic inflammation might further contribute to the development of some malignancies (8), such as gastric mucosa-associated lymphoid tissue (MALT) lymphoma (9–16).

Mucosa-associated lymphoid tissue lymphoma (MALT-L) is a low-grade B-cell lymphoma. It grows slowly and remains confined to one organ for a relatively long time. Stomach MALT-L exemplifies the close link between chronic inflammation and lymphomagenesis. B-cells of MALT-L are related to normal marginal zone cells. Their IgH variable region gene sequences exhibit features of post germinal center B-cells, such as somatic hypermutation (SHM), implying that the clone has expanded in the presence of an antigen (17). MALT-L is often associated with bacterial infection, most commonly by *H. pylori* bacterium (7–9, 15–17).

A possible outcome of low-grade B-cell lymphomas such as MALT-L is the transformation into a more aggressive lymphoma such as diffuse large B-cell lymphoma (DLBCL) (18, 19). Gastric DLBCL is a fast-growing, aggressive B-cell malignancy characterized by diffuse proliferation of large neoplastic lymphoid B-cells (20, 21). DLBCL is known to represent a heterogeneous group of malignancies, comprising either germinal center-like cells exhibiting intra-clonal diversity or "activated B-cell-like" cells, which do not (22, 23).

Abbreviations: AML, acute myeloid leukemia; B-CLL, chronic lymphocytic leukemia; BCR, B-cell receptor; CDR, complementary determining region; CI, confidence intervals; CLN, control lymph node; DLBCL, diffuse large B-cell lymphoma; FL, follicular lymphoma; GHP, gastritis with *H. pylori* background; GNHP, gastritis without *H. pylori* background; *H. pylori*, *Helicobacter pylori*; HTS, high-throughput sequencing; Indels, insertions and/or deletions; MALT, mucosa-associated lymphoid tissue; MALT-L, MALT lymphoma; MID, molecular identification; MM, multiple myeloma; SHM, somatic hypermutation.

During the clonal expansion of B-cells in response to an antigen, Ig gene sequences from clonally related B-cells (i.e., B-cells that are derivatives of the same B-cell ancestor) accumulate mutations via SHM and thus diversify. Clonally related cells are identified by identical V(D)J segments and by highly homologous sequences of the complementary determining region (CDR) 3 of their Ig genes. An easy way to track and analyze the relationships between clonally related Ig gene sequences is by using lineage trees. The tree root is the ancestor sequence, usually the rearranged, pre-mutation sequence. Each tree node represents a single mutation (point mutation, insertion, or deletion). Lineage trees have been used in order to quantify the differences between the dynamics of SHM and antigen-driven selection in different lymphoid tissues, species, and disease situations. Our lineage trees-based mutation analysis has demonstrated its usefulness in previous studies of aging (24), autoimmunity (25–28), and chronic inflammation (29). Recent work on B-cell malignancies done in our lab (30–32) showed differences in tree properties between lymphomas and controls. Lymphoma trees were more branched and had longer trunks compared to controls, indicating a higher intra-clonal diversification and a longer mutational history. Intra-clonal diversification was also shown in chronic lymphocytic leukemia cases (33–35), in marginal zone lymphoma cases (36, 37) and in intestinal DLBCL cases (21). In addition, lymphoma and controls exhibited similar mutation rates and same SHM motifs. Follicular lymphoma (FL), which is considered a less aggressive lymphoma, displayed higher diversity than DLBCL and highest recent diversification events, suggesting that the more aggressive lymphoma diversifies the least (38–40).

In the present study, we used repertoire, lineage tree, and mutation analyses to investigate whether and how B-cell repertoires, clonal diversification, and selection mechanisms in gastritis, gastric MALT-L, and DLBCL differ from each other and from normal responses. The two types of gastritis (positive or negative for *H. pylori*) were found to have similar repertoires and diversification. MALT-L clones were found to be more diversified and had longer mutational histories compared with all other conditions, but the dominant clones of MALT-L (defined as the largest clones in each sample) were different from those of all other conditions. DLBCL dominant clones, however, displayed lower diversification. These results are surprising, as we expected to find similarities between the dominant clones of gastritis and MALT-L and between those of MALT-L and DLBCL, according to the hypothesis that these are often sequential steps of inflammation and transformation.

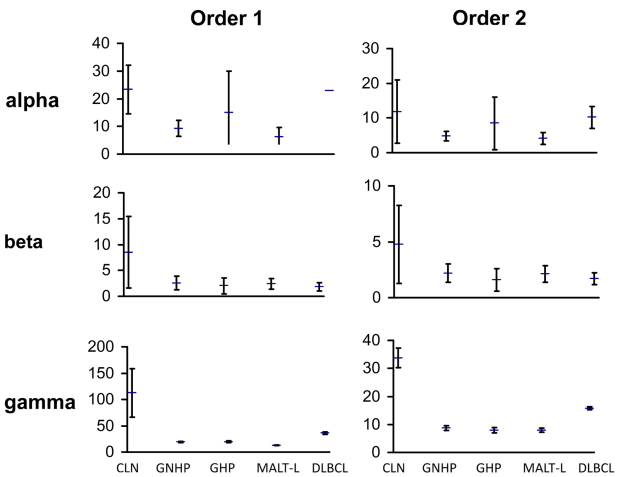


FIGURE 1 | Graphic presentation of alpha, beta, and gamma diversity indices of order 1 and 2. Diversity measures were calculated, as described in the Section "Materials and Methods," based on the Shannon entropy and the Simpson concentration diversity indices. The alpha diversity measure represents the average sample diversity in each condition/population. The gamma diversity measure represents the "global" repertoire diversity across all samples studied in each condition/population. The beta diversity measure represents the diversity component resulting from the variability between samples. In our Ig gene repertoire studies, the abundance data (numbers of unique sequences) of antibody clones in each sample was used to estimate the mean, standard error, and 95% confidence intervals (CI) of the total number of unique sequences in clones within each sample. The CI allows us to compare between diversity indices of different conditions. The error bars show the standard errors.

RESULTS

REPERTOIRES IN GASTRITIS WITH *H. PYLORI* BACKGROUND WERE AS DIVERSE AS THOSE IN GASTRITIS NEGATIVE FOR *H. PYLORI*, AND CONTAINED SIMILAR V(D)J COMBINATIONS

We compared the repertoires in both types of gastritis, with *H. pylori* background (GHP) or without *H. pylori* background (GNHP), and examined the differences between them. We expected the repertoire in GHP to be less diverse due to the response to the bacterium, as previous studies showed that monoclonality is frequently found in GHP samples [(41–43) and others]. In contrast to our expectation, the confidence intervals (CI) of alpha, beta, and gamma diversity indices of both orders were overlapping (**Figure 1**), implying the average individual biopsy diversities, the variability of diversities between individual biopsies, and the overall pool diversities in GHP and GNHP were not statistically different. Indeed, most V(D)J combinations observed were expressed in both gastritis types (**Figures 2A,B**).

Gastritis with *H. pylori* background and GNHP were the most similar conditions (similarity index of 0.543, **Table 1**), although one GHP sample (the second GHP sample in Table S1 in Supplementary Material) had an extremely high alpha diversity index compared to the other two samples (data not shown). This contradicts our expectation of narrower repertoires in GHP samples due to the presence of *H. pylori*. However, if the one highly diverse GHP sample is excluded from the analysis, the confidence interval of alpha of GHP becomes narrower (3.9–11.25), and lower than that of GNHP. It is possible that the highly diverse sample reflected additional ongoing responses.

VH1-3/JH4 was a common combination in both GNHP and GHP VH–JH repertoires, but not so prominent in repertoires of other conditions (**Figures 2A,B**). These combinations contained several DH genes in both GNHP and GHP. However, identification of D genes should be taken with caution, as SoDA always finds a D gene, even when this is based on too-few nucleotides to be reliable. **Table 2** summarizes all common combinations and genes found in our study and their relationships with other clinical conditions as implicated in the literature.

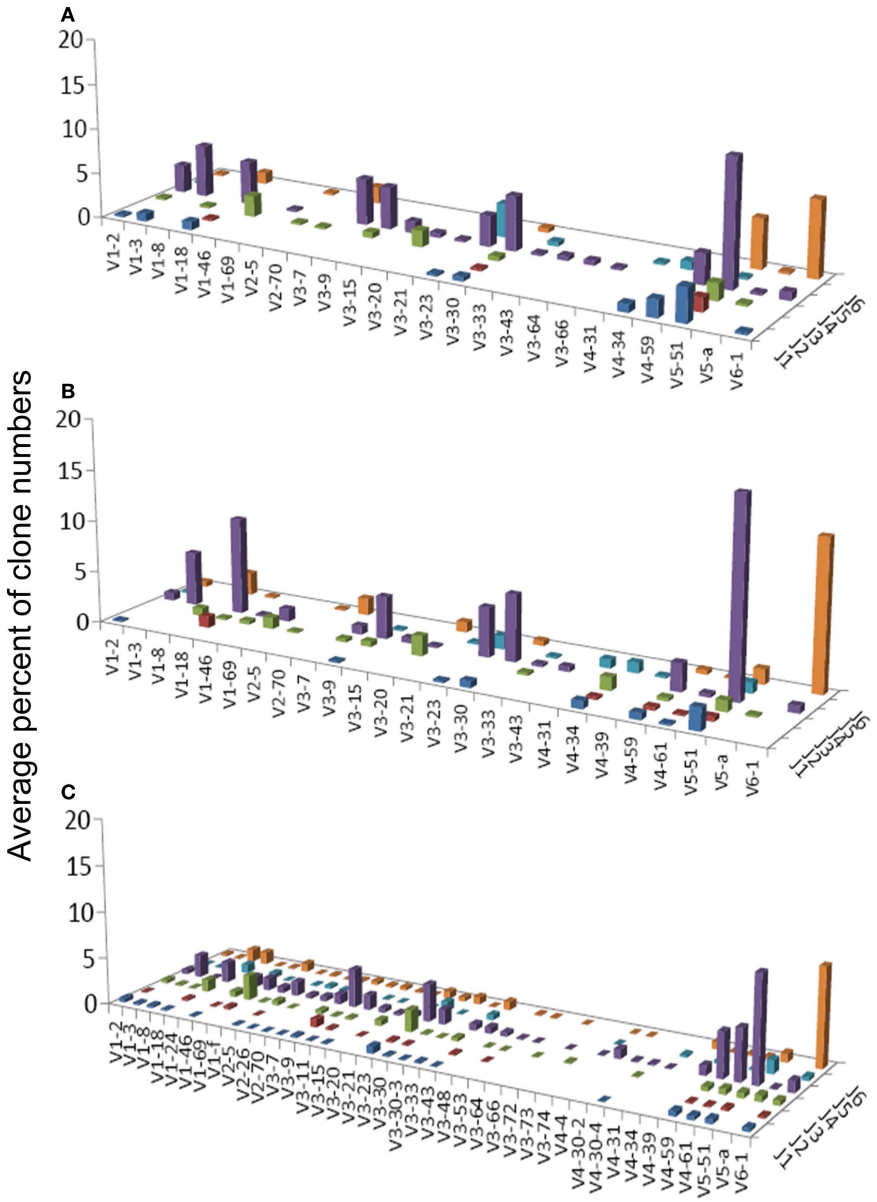


FIGURE 2 | Average percentages of clones in each VH–JH combination, in (A) GHP, (B) GNHP, and (C) CLN samples.

Table 1 | The average similarity between each pair of conditions.

	CLN	GNHP	GHP
CLN	0.211	0.295	0.343
GNHP		0.408	0.543
GHP			0.478

Similarity measures were calculated between all clones in all samples in the compared conditions. In lymphomas, only the dominant clone(s) are relevant, as the rest of the clones in each sample represent other B-cells present in the tissue, which are not related to the malignancy. Thus, MALT-L and DLBCL are not included in the calculation of similarity measures because the dominant clones in these conditions cannot be compared to the full repertoire samples from other conditions.

In the case of GNHP, the DH3 and DH6 families were found to be preferred, as combinations of V6D3J6 and V6D6J6 were used significantly more than expected (**Table 3**; **Figure 2B**). Other over-expressed DH genes were used in less prominent combinations in the observed repertoire. Other combinations used in GNHP and GHP, such as VH3-7, VH3-23, and VH3-30 – all with JH4 – used several DH genes from different DH families.

Gastritis without *H. pylori* background and GHP presented almost identical gene usage patterns, having VH5-51/JH4 and VH6-1/JH6 as the two most frequent combinations (**Figures 2A,B**). VH5-51 and VH6-1 have been shown to often participate in earlier stages of repertoire development via positive selection by auto-antigens (**Table 2**). These two combinations also appeared in our control lymph node (CLN) samples (**Figure 2C**),

Table 2 | A summary of frequent combinations and genes in conditions from our study and from other studies.

Gene or combination	Common in our study in	Appeared in the literature in relation to
VH1-2	MALT-L[a]	Self-reactive antibodies, Bahler et al. (44) Chronic lymphocytic leukemia (B-CLL), primary central nervous system lymphomas, and splenic marginal zone lymphomas, Walsh and Rosenquist (45)
VH1-3	GNHP, GHP	B-CLL, Fais et al. (46)
VH1-18	GNHP, GHP, DLBCL	B-CLL, Fais et al. (46) Autoreactive gene, Yamashita et al. (47) BM-DLBCL, gastric MALT-Ls, Bende et al. (48)
VH1-8	DLBCL	B-CLL, Pimentel et al. (49)
VH1-69	MALT-L[a], DLBCL	Rheumatoid factor, Bende et al. (48), Matsuda et al. (50) Gastric MALT-Ls, Bende et al. (48) B-CLL, Fais et al. (46), Pimentel et al. (49), Johnson et al. (51)
VH2-26/JH5	MALT-L[a]	FL, Bayerl et al. (52)
VH2-26	MALT-L[a]	B-CLL, Pimentel et al. (49) Hairy cell leukemia, Hashimoto et al. (53)
VH3-7	MALT-L Dominant, DLBCL	Rheumatoid factor, Bende et al. (48), Matsuda et al. (50) Rheumatoid arthritis, Nakamura-Kikuoka et al. (54) Sjögren syndrome, Bahler and Swerdlow (55) B-CLL, Fais et al. (46) BM-DLBCL, Yamashita et al. (47) Gastric MALT-Ls, Bende et al. (48)
VH3-23	DLBCL	IgM+ B-cells, Brezinschek et al. (56) Naïve B-cells, Wu et al. (57) Anti-DNA auto-antibodies, Matsuda et al. (50) Hepatitis C virus-related mixed cryoglobulinemia, Perotti et al. (58) Unmutated VH3-23 in transformation from B-CLL into DLBCL, Mao et al. (59) Gastric MALT lymphomagenesis, Sakuma et al. (60), Lenze et al. (61), Alpen et al. (62), Siakantaris et al. (63) BM-DLBCL, Yamashita et al. (47) B-CLL, Pimentel et al. (49)
VH3-30	DLBCL	Rheumatoid factor, Bende et al. (48), Matsuda et al. (50) Gastric MALT lymphomagenesis, Sakuma et al. (60), Lenze et al. (61), Alpen et al. (62), Siakantaris et al. (63) B-CLL, Pimentel et al. (49)
VH3-30/JH4	DLBCL	FL, Bayerl et al. (52)
VH5-51 and VH6-1	CLN, GNHP, GHP, and DLBCL Dominant	Auto-antigens, Matsuda et al. (50)

The dominant segments are marked with "Dominant" indication. The dominant segments/combinations are those that appeared in the largest clone in each sample.
[a]Represents frequent genes or combinations in the repertoire of unique sequences.

so they probably have no specific connection to gastritis or *H. pylori* response. However, the VH1-18/JH4 combination was more frequently used in GNHP than in GHP, and was not prominently observed in other conditions. As VH1-18 is an autoreactive gene and was found in several gastric MALT-Ls (**Table 2**), VH1-18 may be involved in the development of gastritis regardless of the presence of *H. pylori*.

One combination was over-expressed in both types of gastritis (V3D0J5), and two combinations (V6D3J6 and V6D6J6, **Table 3**) were over-expressed in GNHP (as dominant clones) and in DLBCL samples (not the dominant clones). As can be seen in **Figures 2B,C** and **3A**, many sequences used the V6–J6 combination in GNHP and DLBCL, but also in the controls. Thus, this combination is very frequent in immune responses, and cannot be ascribed to a

Table 3 | VDJ combinations that were over-expressed in each condition[a].

Condition	Combination	*p*-Value	Mean difference[b]
GNHP	V2D1J6	0.009	3.57
	V3D5J4	0.002	4.38
	V3D0J5	0.001	7.84
	V6D6J6	0.035	3.84
	V6D3J6	0.042	3.60
GHP	V3D0J5	0.035	10.44
	V5D3J2	0.015	3.26
DLBCL	V1D1J6	0.011	1.30
	Dominant[c]		
	V1D4J3	0.000	4.94
	V4D6J4	0.046	2.70
	V5D1J4	0.001	1.80
	Dominant[d]		
	V5D7J4	0.036	2.20
	V6D6J6	0.000	3.47
	V6D3J6	0.003	3.50

[a]*There were no VDJ combinations that where over-expressed in MALT-L samples, thus MALT-L does not appear in the table.*

[b]*Represents the value of log_2(observed/expected).*

[c]*This combination was found in DLBCL samples number 1, 2, 3 (sample numbers according to Table S1 in Supplementary Material). The dominant combinations are those that appeared in the largest clone in each sample.*

[d]*This combination was found in DLBCL sample number 5.*

specific condition. However, the combination V3D0J5 may represent an antibody that is effective in the gastric environment, related to inflammatory processes, or participates in both.

To conclude, both types of gastritis presented similar repertoires and diversity properties, in contrast to our expectation.

GASTRIC MALT-L EXHIBITED UNIQUE V(D)J COMBINATIONS

Several studies have demonstrated that gastric MALT-L is often associated with a bacterial infection, most commonly by *H. pylori*; another association has been revealed between gastritis and gastric MALT-L (11–13). Therefore, we expected to find similar V(D)J combinations when comparing the two conditions. Surprisingly, the dominant MALT-L V(D)J combinations were very different from those in GHP. While GHP showed an extensive use of JH4 family genes and several common combinations, of which the most frequent was VH5-51/JH4, in MALT-L dominant clones were VH3-7/JH4, VH1-69/JH6, and VH1-2/JH1. VH3-7 is frequently found in rheumatoid factors and was selectively expressed by patients with rheumatoid arthritis and Sjögren syndrome. Preferential use of these genes and combinations has been reported in several types of lymphomas and leukemias (**Table 2**).

Table S2 in Supplementary Material presents the combinations that were over-expressed in one condition while underrepresented in the other. It can be seen (from the "Mean deviation" column) that over-expressed combinations were almost absolutely from either MALT-L or DLBCL samples (dominant clones only). As DLBCL contained dominant combinations that also appeared in other conditions, this supports the observation of different dominant combinations in MALT-L compared to those observed in other conditions. These combinations may relate to the malignancy, but this remains to be explored.

We also compared the dominant clones in DLBCL and MALT-L samples, as in some cases DLBCL appears in association with MALT-L (18, 19). DLBCL is considered in these cases to result from clonal transformation of large cells within the low-grade lymphoma (64, 65). Hence, we expected to identify similar segment combinations on the dominant clones from the two conditions. However, the dominant clones of MALT-L samples were different from those of DLBCL (**Figures 3A,B**). As mentioned above, the dominant clones in MALT-Ls were VH3-7/JH4, VH1-69/JH6, and VH1-2/JH1, while in DLBCLs these combinations were found, but were not the dominant clones. VH5-51/JH4 and VH6-1/JH6 were frequent combinations in all conditions, except in MALT-L, suggesting they may have some advantage in binding common antigens. In terms of unique sequences, MALT-Ls presented completely different dominant combinations from DLBCL (VH1-2/JH1, VH1-69/JH6, VH2-26/JH5, and VH3-7/JH4, data not shown). Preferential use of these genes and combinations has been reported in several types of lymphomas and leukemias (**Table 2**). In addition, VH3-7/JH4, which was the most frequent combination in MALT-L dominant clones, appeared in all other conditions but with dramatically lower numbers. As mentioned above, VH3-7 participates in the formation of autoantibodies and was found in several gastric MALT-Ls. Some MALT-Ls were found to use VH genes previously associated with auto-reactivity. This suggests that B-cells in MALT-L react with self-antigens (66), different from those that arouse in GHP and DLBCL responses.

MALT-L DOMINANT CLONES HAD LONGER DIVERSIFICATION HISTORY, IN CONTRAST TO DLBCL CLONES

Lineage trees of the MALT-L dominant clones had significantly longer trunks (T) and path lengths (PLmin), which are tree length measures, than all other conditions (**Figure 4**; Figure S1 in Supplementary Material). In addition, according to the correlation of tree properties with the dynamic parameters of the secondary B-cell response (67), longer trunks correlate with a lower initial affinity, and longer paths also correlate with a lower selection threshold. This suggests that diversification history in MALT-L dominant clones was longer than that of other conditions.

On the contrary, DLBCL dominant clones had significantly shorter trunks and path lengths than those of GHP (and MALT-L), and in general, the lowest tree length measures. Dominant clones of DLBCL presented similar tree length measures to those of CLN (**Figure 4**; Figure S1 in Supplementary Material). This is in line with the above-described observation of similarity between DLBCL and CLN. The shorter lengths observed in DLBCL, which correlate with high initial affinity and selection threshold, may indicate a shorter diversification process compared to MALT-L and GHP (21). A possible explanation for this is that, because MALT-L

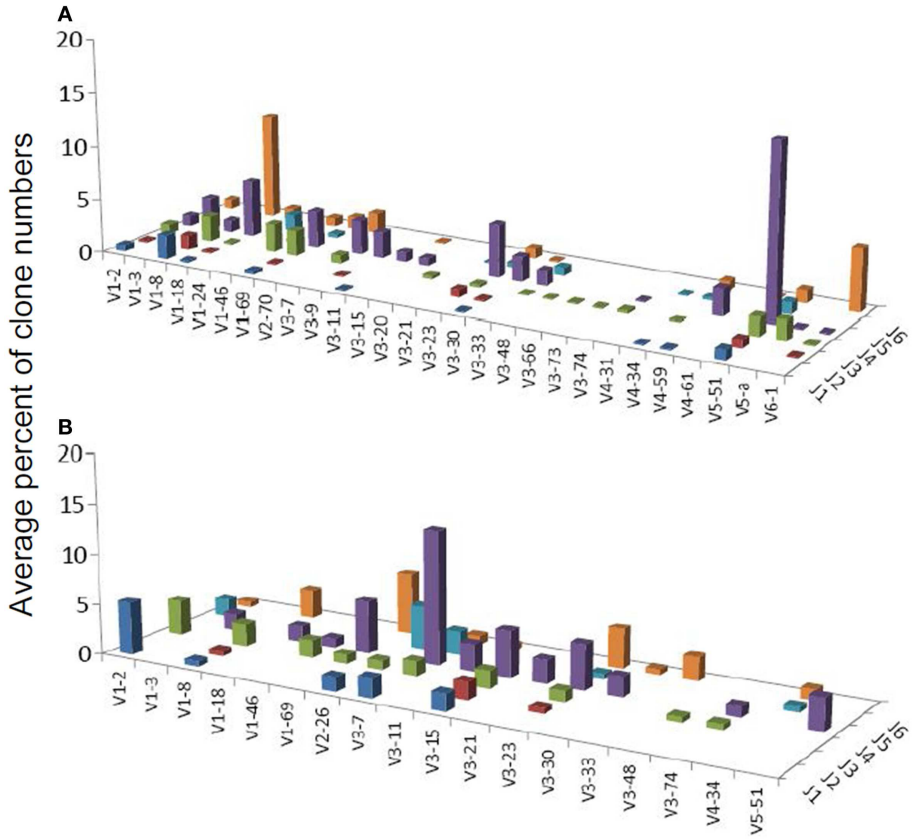


FIGURE 3 | Average percentages of clones in each VH–JH combination, in (A) DLBCL and (B) MALT-L samples.

is an indolent lymphoma and DLBCL is an aggressive lymphoma, the latter usually has less time to diversify until it is discovered and treated. Figures S2 and S3 in Supplementary Material show representative examples for MALT-L and DLBCL dominant trees, respectively.

The fact that MALT-L dominant clones had larger trunks, path lengths, and distance from the root to any split node, thus probably lower initial affinity than DLBCL dominant clones, suggests that DLBCL dominant clones started as responses to specific (yet-unknown) antigens, with probably higher initial affinity than the responses that initiated MALT-Ls. That is, high affinity and vigorous response may be risk factors for aggressive lymphoma development. In terms of selection, these results show that selection thresholds in MALT-L dominant clones were the lowest among all other conditions. Low selection pressure may simply be the result of abundance of antigen, and this may indeed be the case in gastric MALT-Ls.

Dominant clones from the two types of gastritis presented similar tree length measures, which correlate with the observed similar repertoires.

DISCUSSION

In this study, we investigated the relationships between four related conditions of the stomach: gastritis positive or negative for *H. pylori*, gastric MALT-L, and gastric DLBCL. As previous studies showed, these conditions sometimes appear successively, as prolonged stimulation during chronic gastritis may result in the development of gastric MALT-L, which in some cases further transforms into DLBCL. We examined the clonal repertoires of the IgH variable region genes (or in the case of lymphomas, the dominant clones, which are defined as the largest clone in each sample) and the lineage tree characteristics in each condition, in order to find similarities or differences between these conditions.

Both types of gastritis presented similar IgH variable region gene repertoires and lineage tree characteristics, in contrast to our pre-study assumptions. However, although the GNHP biopsies were negative for *H. pylori*, it could be present in the tissue in undetectable amounts, and thus affect the repertoire of the B-cells in its surroundings. Moreover, both types of gastritis used the VH1-18 gene, which may be involved in the development of gastritis regardless of the presence of *H. pylori*; this remains to be elucidated, as this was not the dominant combination in both type of gastritis. We expected the repertoire in GHP to be less diverse than that of GNHP due to the response to the bacterium, which is expected to elicit only specific clones. In contrast to our expectations, GHP samples showed at least as diverse repertoires as GNHP (**Figures 1** and **2A,B**). An explanation for a high diversity in GHP might be the phase variation of *H. pylori*, which is the generation of intra-strain diversity that is important for bacterial

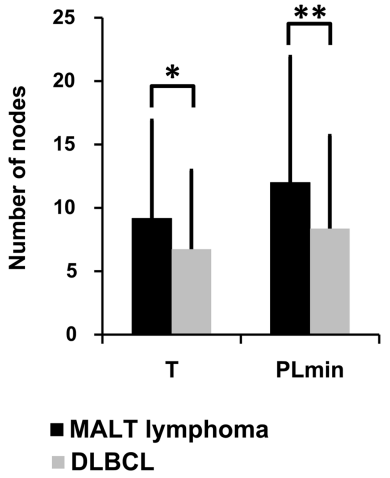


FIGURE 4 | Lineage tree analysis – comparison between dominant clones from the three MALT-L samples (24 trees) and the five DLBCL samples (47 trees). There was more than one tree per sample, as we included all clones with the same VH and JH genes (but different alleles) in the dominant clone in each sample, because they might be related to the dominant clone and falsely attributed to other alleles. Significant differences were found in trunk length (T) and minimal path length (PLmin). An asterisk (*) represents *p*-value <0.01; two asterisks (**) represents *p*-value <0.005.

niche adaptation (68), and could cause variability not only in the bacterial strains but also in the responding antibody repertoire. Hussell et al. (69, 70) showed that the extreme variability of *H. pylori* strains led to diverse T cell responses. Moreover, they showed that B-cells did not respond to *H. pylori* themselves, but required contact-dependent help from *H. pylori*-specific T cells, and their Ig genes responded to auto-antigens, similar to our observations. Alternatively, the repertoires in the biopsies – even GHP biopsies – may reflect immune responses to a variety of pathogens, including but not limited to *H. pylori*.

Although in many cases DLBCL is associated with MALT-L, the two types of lymphomas presented different dominant clone combinations and lineage tree characteristics. DLBCL may develop after prolonged stimulation during gastritis, derive from a low-grade malignant clone, or it can initiate *de novo*, depending on the mutations in each clone. In this study, as MALT-L and DLBCL presented different dominant clone combinations, in contrast to our expectations; we speculate that in these cases DLBCL may have initiated *de novo*. MALT-L samples presented different lineage tree characteristics from those of all other conditions, although we expected MALT-L to resemble GHP. In fact, we identified preferential use of the autoreactive gene VH3-7 in MALT-L samples. VH3-7 was one of the common VH genes in GHP (but not the dominant clone). These findings suggest that gastric MALT-L is derived from highly restricted B-cell subsets probably resulting from specific antigenic stimulation, such as with *H. pylori* (15). It is possible that B-cells in MALT-L react with self-antigens (66), however, the role of self-antigens in the development of the malignancy has yet to be examined. Moreover, lineage tree drawings demonstrated longer trunks and path lengths in MALT-L dominant clones, compared with all other conditions. These differences in tree characteristics correlate with lower initial affinity and lower selection threshold, respectively. Low selection pressure may simply be the result of abundance of antigen, and this may indeed be the case in gastric MALT-L. The above may indicate that MALT-L has undergone a longer mutational history than other conditions. On the contrary, the shorter lengths observed in DLBCL dominant clones may be a result of shorter diversification and responses to specific (yet-unknown) antigens, with higher initial affinity compared to MALT-L and the two types of gastritis (21). The latter may be risk factors for aggressive lymphoma development.

We observed some similar VH–JH combinations in all conditions, together with over-expressed and preferred combinations unique to gastritis, MALT-L, and DLBCL samples. These combinations should be investigated in order to further understand their role in the development of each condition. For example, the relatively extensive use of combinations, which were previously found in other malignancies, in DLBCL samples in this study reinforces the notion that some DLBCL clones had developed from MALT-L clones. Moreover, the fact that the dominant combinations were identical in all conditions except in MALT-L (and that MALT-L did not contain this combination at all) is interesting and should be further investigated. We also identified frequent VH genes in GHP repertoires and in the dominant clones in MALT-L samples, which were also found in autoimmune diseases. It was previously shown that some autoreactive B and T cells are activated during *H. pylori* infection (71). The connection between the appearances of these VH genes in GHP and MALT-L samples from our study and in autoimmune diseases remains to be explored. There was no prominent trend toward any V, D, or J gene family in the over-expressed combinations in each of the conditions. However, combinations that were over-expressed in lymphoma dominant clones compared to other conditions had a clear preference toward the use of VH2, JH1 in MALT-L and VH1, JH2 in DLBCL. The role of these combinations and gene families in lymphomas is still unknown.

It should be noted that, because we studied formalin-fixed paraffin-embedded archival biopsies, we had access to only a limited number of biopsies, and limited amounts of DNA, as DNA in the preserved biopsies is often denatured (72). More samples in more conditions will have to be studied in order to give a clearer picture of the roles of specific V(D)J genes and combinations in inflammation and malignancies. Moreover, the similarity between conditions is also affected by the limited number of samples in each condition. This may cause the similarity calculations to be biased toward random features of the samples that may characterize a certain condition, and thus affect the interpretations. However, the similarity between samples within each condition was not very large, so it is unlikely that some random feature is common to all samples of the same condition. Coincident with obtaining more samples, we intend to use the Illumina high-throughput sequencing (HTS) in future studies, in order to avoid the 454 sequencing artifacts, which resulted in discarding many sequences (73). A possible argument could be raised regarding the exclusion of duplicate sequences from our analysis. As mentioned in the Section "Materials and Methods," duplicate sequences may stem from the PCR amplification, and may cause misidentification of dominant clones. Although we stringently defined duplicate sequences as those that had the exact nucleotide composition

and equal lengths, most of the sequences differed not only by their lengths, but also in the mutations they include. Table S5 in Supplementary Material summarizes the distributions of the number of sequences in the tree-nodes in the dominant and second dominant clones in lymphoma samples. As can be seen from Table S5 in Supplementary Material, most of the nodes contained only one sequence, implying that most of the sequences differed not only in their lengths, but also in the mutations they include. There were, however, several nodes that contained more than one sequence. This does not necessarily indicate that those sequences are duplicates, as the clone-tree is built according to the aligned sub-sequences that are shared between all sequences in the clone. This means that sequences included in a specific node could differ in their ends, and not necessarily be duplicates.

All conditions in this study presented similar mutation rates and the same SHM targeting motifs. Replacement and silent mutation analysis revealed a strong selection against replacement mutations in the CDR regions of all conditions. This may indicate that most of the examined IgH variable region gene sequences represented B-cell receptors (BCRs) that were already highly specific to their antigens and thus selection operated against replacement mutations in their CDR regions, which are responsible for antigen binding. These results are also consistent with previous work on lymphomas from our lab (30).

In this study, we used the Morisita similarity index in order to measure similarity between conditions (as mentioned in the Section "Materials and Methods"). Similarity measures were calculated between all clones in all samples in the compared conditions. In lymphomas, only the dominant clone(s) are relevant, as the rest of the clones in each sample represent other B-cells that are also present in the tissue, but are not related to the malignancy. Thus, the MALT-L and DLBCL conditions were not included in calculation of similarity measures exactly because the dominant clones in these conditions cannot be compared to the whole other samples in other conditions. In addition, Morisita similarity index is of order 2, which emphasizes large clones. This affects the similarity results. However, we are interested in the larger clones in each condition as they probably represent the dominant responses.

In summary, we showed that gastritis positive or negative for *H. pylori* presented very similar IgH variable region gene repertoires. This suggests that the diverse stomach repertories do not change much due to the presence of the bacteria, and moreover, GHP does not become oligoclonal (or at least with narrower repertoire) due to *H. pylori*. MALT-L, however, presented different and unique dominant clone gene combinations, which can result from specific antigenic stimulation. As was mentioned in the Section "Introduction," several studies showed that *H. pylori* causes gastritis, and suggested that prolonged gastritis can lead to MALT-L, and that prolonged MALT-L can develop into DLBCL. This flow and graduation of diseases led us to the assumption that the repertoires (VDJ combinations) in these conditions would be similar, because these conditions were initiated by the bacteria, and several clones got out of control to progress into lymphoma. Moreover, the diversity in these conditions was expected to be narrower than that in CLN, and to be progressively lower as the conditions proceed toward the aggressive lymphoma. However, the results differed from what we expected. In addition, some combinations did appear in several conditions, but not in MALT-L, and the DLBCL dominant clones also appear in other conditions (so they were not unique to the cancer). We speculate that the transformation into MALT-L, after the prolonged stimulation by the chronic GHP, amplified specific combination(s) that were also found in GHP but in a lower frequency (such as VH3-7). The two types of lymphomas differed in their dominant clone gene combinations and lineage tree characteristics, suggesting differences in the abundance of antigens, if not in their nature, which remain to be explored.

MATERIALS AND METHODS

HISTOPATHOLOGICAL SPECIMENS

Five gastric DLBCL biopsies, 3 gastric MALT-L biopsies, 10 chronic gastritis biopsies (3 with *H. pylori* background and 7 that were negative for *H. pylori*), and 19 reactive lymph node biopsies (which served as controls), each from a different patient, were selected from the pathology department archives in the Sheba Medical Center (Table S1 in Supplementary Material). Tissue biopsies were taken during resection procedures and were used in this study in accordance with institutional Helsinki committee guidelines and approval. Histochemical stains, by Hematoxylin-Eosin (H&E) and Giemsa, were performed for histological evaluation and *H. pylori* identification. For diagnosis of lymphoproliferative disease and characterization of lymphocyte populations, immunohistochemical stains (e.g., CD20, CD3, CD23, CD21, cyclin-D1, Ki67, and IgD) were also performed. All cases were revised by two independent pathologists to confirm the diagnosis.

DNA EXTRACTION

Paraffin-embedded blocks were cut using a microtome to get extremely thin slices of tissues (sections). Each of the biopsies was consecutively cut to yield 10–20 sections of 4 μ each, depending on tissue size. All sections from each biopsy were inserted into an eppendorf tube with 200 μl water (Sigma) and were heated in 90°C until the paraffin was melted. After the tubes were centrifuged at full speed (14000 rpm) for 1 min, a paraffin ring was created and could easily be removed from each of the tubes. Water was drawn from the tubes while tissues remained in the tubes. Extraction of DNA was then performed using the QIAamp DNA Mini Kit (or the QIAamp DNA Micro Kit for very small samples) according to the QIAGEN protocol.

In some cases, a micro-dissection was needed. First, H&E stained thin sections were reviewed by a pathologist, and areas of interest were outlined. The tissues were cut from 5 to 10 sections that were placed onto five slides per tissue. The slides were heated in 90°C for 15 min. Next, the slides were hydrated (5 min soaking in Xylene, brief immersion in Ethanol 100, 96, and 70% in this order). The slides were then placed to dry. According to the outlined stained slides of each tissue, the hydrated slides were scratched with buffer ATL (QIAamp DNA Mini Kit, QIAGEN) and sample scrapings were picked up into eppendorf tubes. Extraction of DNA was then performed using the QIAamp DNA Mini Kit (or the QIAamp DNA Micro Kit for very small samples) according to the QIAGEN protocol.

PCR AMPLIFICATION AND HIGH-THROUGHPUT SEQUENCING

For each sample taken from each biopsy, semi-nested PCR was performed using the same forward primers for the two PCR rounds and two different reverse primers as described below.

Forward primers from FR2 region:

VH1: 5′-TGCGMCAGGCCCCYGGACAAR-3′,
VH2: 5′-ARGRAAGGCCCTGGAGTGG-3′,
VH3: 5′-CCAGGCTCCAGGSAAG-3′,
VH4: 5′-MGGGAAGGGRCTGGAGTGG-3′,
VH5: 5′-GAAAGGCCTGGAGTGGATGGG-3′,
VH6: 5′-TTGAGTGGCTGGGRAGGAC-3′.

Reverse primers:

First round – JH1R: 5′-TGAGGAGACGGTGACCAGGGT-3′,
Second round – JH2R: 5′-TGACCRKGGTHCCYTGGCCC-3′.

There is no specific primer for VH7 gene family in the FR2 region, as the VH7 primer in this region is very similar to that of VH1, thus, it amplifies the VH1 family and creates a very strong VH1 bias.

The primers were augmented for HTS experiments by the addition of 5′ sequencing adapter elements and 10-nucleotide unique sample molecular identification (MID) tags according to the 454 FLX Titanium chemistry protocol (Roche) (74). Proofreading Taq DNA polymerase (ABgene) was used in PCR reactions according to the manufacturer's protocol. PCR reaction was performed on a sample of 50 ng DNA from each sample, with slight changes according to calibration (because DNA was taken from different tissues, and each tissue can differ in the percentages of lymphocytes). PCR products were separated on a 2% agarose gel stained by ethidium bromide. Clear bands were cut from the gel and DNA was extracted using the MinElute Gel Extraction kit (QIAGEN), according to the manufacturer's protocol. Sequencing of small samples of the PCR products by the classic Sanger method (after cloning to pGEM – T-easy vector) was performed in order to make sure they are Ig gene amplifications and the sequences of the primers and the tags are intact. DNA concentration of PCR products from each sample were determined by PicoGreen dye and fluorospectrometer (Nanodrop). According to these results, a mixture containing 10^9 molecules of PCR products from each sample was prepared and sent to sequencing. HTS was performed using the 454 GS FLX Titanium platform by DYN Diagnostics Ltd., the sole representative in Israel of Roche Diagnostics. Raw data files containing a total of ~120,000 reads were received when the HTS was completed. Raw data files can be downloaded from the SRA database, accession number PRJNA206548 (Runs: SRR873440, SRR873441, SRR873442).

HTS RAW DATA PRE-PROCESSING

To process the 454 raw data, we used our program Ig-HTS-cleaner (73). Ig-HTS-cleaner discards artifact sequences, assigns the sequences to samples according to their MID tags, identifies primers, and discards sequences much shorter or longer than the expected length of an Ig variable region gene, or sequences with average quality scores below a defined threshold. Parameters used in the Ig-HTS-Cleaner run were as follows. Average quality score threshold of 20, a maximum of 2 allowed mismatches in the primer search, 75% of the primer's length to search, and a range of 25 bases at the ends of the read for the MID and primers search (Table S3 in Supplementary Material).

Next, we discarded duplicate sequences, which are completely identical sequences, from each sample. We cannot exclude the possibility that duplicate sequences are a result of the PCR amplification; hence the existence of many identical sequences in a sample does not necessarily indicate that the sequence is found in the original biopsy in the same frequency.

Afterward, we used our program Ig-Indel-Identifier (Ig Insertion – Deletion Identifier) (73), in order to identify legitimate and artifact insertions and/or deletions (indels) in the sequences. Parameters used in the Ig-Indel-Identifier run were as follows. HPT length was set to 2, quality score threshold (for suspected point mutations) of 25, and the number of sequences in the same clone containing the same indel was set to 1 (Table S4 in Supplementary Material). **Table 4** presents the numbers of unique sequences from each condition after discarding sequences with suspected indels. These were the final numbers of sequences that were analyzed.

GERMLINE VDJ SEGMENT IDENTIFICATION AND ASSIGNMENT INTO CLONES

Clonally related sequences were identified by identical V(D)J segments and by highly homologous sequences of the CDR3 of their Ig genes. For gene segment identification, we used SoDA (75). We computationally grouped the sequences into clones based only on their V, D, and J segments. We aligned clonally related sequences using ClustalW2 (76), in order to confirm that the CDR3 in the clonally related sequences were highly homologous. If not, we separated the sequences into clonally related groups according to the different CDR3 sequences.

REPERTOIRE ANALYSIS

We enumerated the clones based on V(D)J combinations. Results are presented as the average sample percentages of clones of each VH–JH combination, across all samples within the same group. Using the percentages normalizes for different numbers of sequences and/or clones, due to sampling of different numbers of B-cells or obtaining different DNA quantities in each case.

In order to examine the relationships between the VDJ combinations used in each repertoire, we needed to compare the observed repertoires to repertoires predicted under some model, for example, under the assumption that the expression of each gene in each VDJ combination (e.g., V1D1J1) is independent of that of other genes. Immunologists call this assumption "the product rule" (77). Deviations from this assumption can thus point at interdependencies between the V, D, and J genes. We decided to look only at the gene family level, as higher resolution (genes, alleles) would give extremely large numbers of possible combinations, far from the number of combinations observed and therefore the frequencies of each expected combination at the gene or allele level would be close to zero. Thus, each observed gene combination would be significantly different from the expected. Using only families of the VDJ segments would solve this problem. For each sample, we counted the number of unique sequences that used each V, D, or J family. We then calculated the frequency of each V/D/J family as the number of unique sequences using this family divided by the total number of unique sequences in the sample. We then created all possible

Table 4 | Number of unique sequences[a], without suspected indels[b], from each condition.

	CLN	GNHP	GHP	MALT-L	DLBCL	Total
Number of patients (samples)	19	7	3	3	5	37
Number of unique sequences	23,308	4,676	3,373	3,851	4,389	39,597
Range of sequences[c]	384–3,353	75–1,105	513–1,406	838–1,854	267–1,601	
Dominant clone sizes[d]				360 (249)	162 (49)	
				461 (399)	325 (257)	
				408 (321)	58 (42)	
					78 (47)	
					418 (193)	

[a] *Unique sequence: a sequence that differs from all other sequences due to one or more insertion(s), deletion(s), or point mutation(s).*

[b] *After sequences with suspected indels were discarded.*

[c] *The lowest-to-highest numbers of unique sequences without suspected indels in each sample from each condition.*

[d] *The number of unique sequences in the dominant clone, and the number of the second dominant clone (in parentheses), in the lymphoma samples. Each couple of numbers represents one-sample.*

combinations that can be made using the observed VDJ families, and defined their expected frequencies as the product of the V/D/J family frequencies calculated in the previous step. Next, we calculated the actual frequencies of the observed combinations (number of unique sequences in each observed combination divided by the total number of unique sequences). There was no point in creating combinations with families that did not appear in the sample, as there was no meaning of calculating frequencies of non-existing combinations.

In order to know whether a combination was expressed more or less than expected, we calculated the expression: $\log_2$(observed/expected). If a specific combination was over-expressed compared to the expected frequency, the ratio inside the logarithm would be larger than 1, as the observed frequency would be larger than the expected, and thus the logarithm would be positive. On the other hand, if a specific combination was under-expressed compared to the expected frequency, the ratio inside the logarithm would be smaller than 1, and thus the logarithm will be negative. Combinations that were not observed at all received the value $(-\infty)$, because the expression inside the logarithm was zero. This step was repeated for each sample. It is important to note that most of the combinations (>80%) were observed in a significantly different frequency than expected. Out of these combinations, 99% were under-represented (because the number of combinations observed is smaller than the potential number), and only 1% of the combinations were over-expressed compared to the expected frequency, and any such case of over-expression was thus particularly noticeable.

The final step was to unite all combinations from all samples as follows: we created a matrix, where rows represented samples and columns represented VDJ combinations. For each combination and for each sample, we inserted the logarithm that was calculated as above. If a sample did not have a specific combination, the cell would be left unfilled. The full matrix was used to carry out the statistical tests. In order to examine whether some combinations tend to appear more or less than expected, we carried out a one-sample t-test on each of the conditions. In order to examine differences in combination usage between conditions, we carried out a two-sample ANOVA test.

In order to graphically present repertoires, we only plotted V–J repertoires, not showing the DH segments used in each VH–JH combination. There are several ways of presenting also the DH genes used in the repertoires (78, 79). However, as mentioned above, DH segments are sometimes misidentified, so we preferred to focus on V and J segments.

DIVERSITY ANALYSIS

Clones in samples can be regarded as species in habitats

In the case of lymphocyte clonal repertoire samples (e.g., those obtained from tissue biopsies), we treat each sample as a sample from a habitat, in which the "species" are the BCR or TCR clones found in the sample. Each of the clones may be composed of a number of different sequences. In TCR clones, all sequences are identical, but in BCR clones sequences from the same clone may be different due to SHM, and one may choose to use only unique sequences found, or all sequences including multiplicate ones. The latter choice depends on whether identical sequences coming from different cells can be identified as such, or cannot be distinguished from sequence duplications caused by PCR amplification. If the former is true (as when using random barcoding in the PCR primers), then the number of sequences that come from different cells is known, and can be used to estimate diversity. If not, then TCR diversity cannot be estimated, and BCR diversity can only be estimated based on the numbers of unique sequences and thus would usually only gives a minimum estimate of the total diversity, as we have done in this study.

Diversity indices

In order to quantify the diversity of clonal repertoires (such as antibody/BCR or TCR gene repertoires) in each experimental or clinical condition, and later to be able to compare between two or more conditions, we used diversity indices (such as the Species Richness, the Shannon entropy, or the Simpson concentration, which are indices of order 0, 1, and 2, respectively) (80). These indices take into account the number of species and (in indices of order >0) the frequency of members of a species (in our case, sequences) of each species (in our case, clone) in each habitat sample. In indices of order 0, diversity is defined simply as number of

species (in our case, lymphocyte clones) in a sample. In order 1 indices, clone size (or frequency) is taken into account, as in the Shannon entropy when diversity is the sum of $\left[-p_i^* \ln\left(p_i\right)\right]$, where *i* represents a species or clone and p_i represents its size (the number of members/sequences, see below). Order 2 indices attribute more weight to large clones, as in the Simpson concentration, which is the sum of p_i^2. In our studies, we used both order 1 and 2 diversity indices, i.e., the Shannon entropy and the Simpson concentration.

Diversity measures

From the sample diversity indices, we have calculated the alpha, beta, and gamma diversity measures for each condition (80). The alpha diversity measure represents the average sample diversity in each condition/population. In order to calculate alpha, we calculate the alpha diversity of each sample, and then average over all samples from the same condition. The gamma diversity measure represents the "global" repertoire diversity across all samples studied in each condition/population. It is calculated as the diversity of the pool containing all the sequences from all the samples from the same condition/population.

Finally, the beta diversity measure, which represents the diversity component resulting from the variability between samples, is derived from the alpha and gamma measures using the method of Jost et al. (80). The beta diversity measure is calculated as the gamma diversity of each condition/population divided by the alpha diversity (average of the diversities of individual samples). In order to allow an intuitive comparison between the diversities of each of the groups, all the diversity measures can be expressed as their number equivalents (80), which reflect the number of equally sized clones needed to produce the given value of the diversity index.

Estimating the full repertoire from the sample

Considering the large number of sequences that were observed only once in each sample, it is likely that many rare clones in an individual's original full repertoire were not detected. To account for the presence of unobserved "species" (clones), all diversity measures can be estimated for whole repertoires (rather than calculated for the sample) using the method described by Chao and Shen (81), which is based on a non-parametric estimation of diversity indices where there are undetected species. Chao and Shen's approach utilizes the concept of sample coverage to adjust the diversity indices for clones that escaped sampling. The sample coverage is estimated from the proportion of species/sequences that are observed only once within a sample.

In our Ig gene repertoire studies, the abundance data (numbers of unique sequences) of antibody clones in each sample were used to estimate the mean, standard error, and 95% CI of the total number of unique sequences in clones within each sample. This was done using SPADE©, a program designed for diversity calculations (81). The alpha diversity for each sample, and the gamma diversities for combined samples in each condition, were then calculated from the order 1 or 2 diversity indices of the estimated total repertoires, also using SPADE©(81). In principle, beta is calculated as the average alpha of all samples in the condition divided by the gamma of the condition, as explained above. In order to compare between conditions, however, we needed to calculate CI for beta. This was done by calculating beta index per sample (alpha of the sample divided by the gamma of the condition) and then calculating the CI for each condition (**Figure 5**).

SIMILARITY ANALYSIS

Another method we used to compare between conditions is the Morisita similarity index (82). SPADE©(81) was used to calculate a similarity matrix, in which we measured each individual repertoire's similarity to all other individual repertoires. The average of similarity indices of individuals in a given group to those in another group represents the similarity index for the comparison between the two groups. A value close to 1 represents high similarity between two groups, and a value close to 0 represents low similarity.

The highest values of the Morisita similarity indices representing the highest similarity were rather low and relatively far from 1, indicating the sensitivity of this method. However, they were consistent with observed repertoire diversities.

Ig LINEAGE TREE ANALYSES

Clonally related Ig gene sequences from each sample were used to create mutational lineage trees using our program IgTree©(83), as described in previous work (29, 30). All trees were measured using our program MTree©, quantifying the graphical properties of the trees (84, 85). A thorough statistical analysis has concluded that seven specific tree characteristics possess the highest correlation values with the biological parameters and are hence most informative (67). As described there, these properties are the minimum root to leaf path length, the average distance from a leaf to the first split node/fork, the average outgoing degree, that is the average number of branches coming out of any node, the root's outgoing

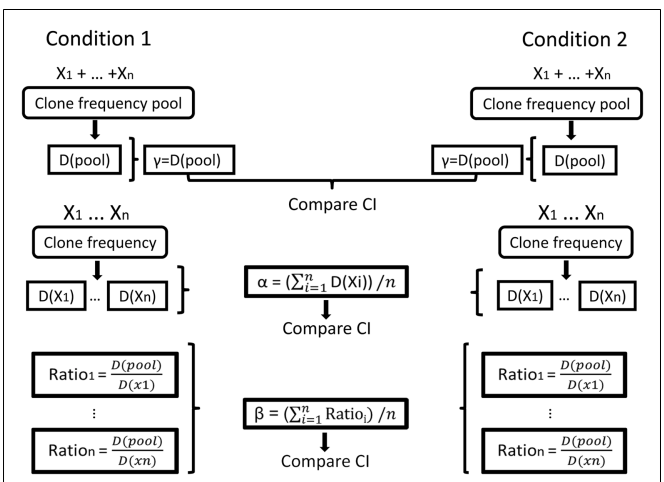


FIGURE 5 | Illustration of the calculation of repertoire diversity. First, the diversity indices (Shannon entropy, Simpson concentration index, etc.) are calculated for the pool of samples in each condition and also for each sample separately (samples are denoted by $X_1, \ldots, X_n$. Diversity indices are denoted by D (sample) or D (pool)). Next, the distribution measures (α, β, γ) are calculated for each condition using the samples. γ is the pool diversity, α is the mean sample diversity, and β for each sample is γ divided by D (sample). Finally, the CIs of the distribution measures are compared between the populations/conditions.

degree, minimum distance between adjacent split nodes/forks, the length of the tree's trunk and minimum distance from the root to any split node/fork. The analysis in this study has thus focused on these properties. Comparison between lineage tree characteristics of different conditions was done using the non-parametric Mann–Whitney *U*-test, as normal distributions (required by tests such as Student's *t*-test) could not be assumed. We used the FDR correction (86) for multiple comparisons.

Replacement (R) and silent (S) mutation analysis methods attempt to measure the extent of selection operating on the diversifying clones. These methods compare the frequencies of replacement mutations found in the frame-work and CDR regions of mutated Ig gene sequences to their expected frequency, based on codon usage of the germline sequence. We used the updated focused binomial test by Hershberg et al. (87, 88). The numbers of observed mutations were pooled for each data group by IgTree©, and the new focused binomial formula (88) was calculated using Microsoft Excel©. This measure was also performed for each sample separately, yielding the same results; however, when comparing conditions, we chose to show the pooled analysis for simplicity. Additional mutational analyses were carried out as described in previous studies (27, 28, 30), however, no significant differences between the conditions were found.

AUTHOR CONTRIBUTIONS

Miri Michaeli performed all steps from DNA extraction from samples to bioinformatical analysis of the sequences, and wrote the manuscript. Hilla Tabibian-Keissar supervised the molecular process. Ginette Schiby revised the samples to confirm the diagnosis. Gitit Shahaf and Yishai Pickman developed the diversity analysis. Lena Hazanov developed the bioinformatical analyses. Kinneret Rosenblatt was in charge of the laboratory in which the molecular work was performed. Deborah K. Dunn-Walters advised the author throughout the study. Ramit Mehr and Iris Barshack supervised the molecular work and the analyses performed, and finalized the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported in parts by an Israel Science Foundation (grant number 270/09, to Ramit Mehr); and a Human Frontiers Science Program Research Grant (to Ramit Mehr and Deborah K. Dunn-Walters). The work was part of Miri Michaeli's studies toward a combined M.Sc/Ph.D. degree in Bar-Ilan University, and she was supported by a Combined Technologies M.Sc Scholarship from the Israeli Council for Higher Education.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00264/abstract

REFERENCES

1. Schubert TT, Schubert AB, MA CK. Symptoms, gastritis, and *Helicobacter pylori* in patients referred for endoscopy. *Gastrointest Endosc* (1992) **38**:357–60. doi:10.1016/S0016-5107(92)70432-5
2. Kusters JG, van Vliet AHM, Kuipers EJ. Pathogenesis of *Helicobacter pylori* infection. *Clin Microbiol Rev* (2006) **19**:449–90. doi:10.1128/CMR.00054-05
3. Sipponen P, Hyuarinen H. Role of *Helicobacter pylori* in the pathogenesis of gastritis, peptic ulcer and gastric cancer. *Scand J Gastroenterol* (1993) **28**:3–6. doi:10.3109/00365529309098333
4. Veldhuyzen van Zanten SJ, Sherman PM. *Helicobacter pylori* infection as a cause of gastritis, duodenal ulcer, gastric cancer and nonulcer dyspepsia: a systematic overview. *Can Med Assoc J* (1994) **150**:177–85.
5. Nordenstedt H, Graham DY, Kramer JR, Rugge M, Verstovsek G, Fitzgerald S, et al. *Helicobacter pylori*-negative gastritis: prevalence and risk factors. *Am J Gastroenterol* (2013) **108**:65–71. doi:10.1038/ajg.2012.372
6. Ryan JL, Shen Y-J, Morgan DR, Thorne LB, Kenney SC, Dominguez RL, et al. Epstein-Barr virus infection is common in inflamed gastrointestinal mucosa. *Dig Dis Sci* (2012) **57**:1887–98. doi:10.1007/s10620-012-2116-5
7. Karlsson FA, Burman P, Loof L, Mardh S. Major parietal cell antigen in autoimmune gastritis with pernicious anemia is the acid-producing H+,K+-adenosine triphosphatase of the stomach. *J Clin Invest* (1988) **81**:475–9. doi:10.1172/JCI113344
8. Sipponen P, Kosunen TU, Valle J, Riihelä M, Seppälä K. *Helicobacter pylori* infection and chronic gastritis in gastric cancer. *J Clin Pathol* (1992) **45**:319–23. doi:10.1136/jcp.45.4.319
9. Segal ED, Cha J, Lo J, Falkow S, Tompkins LS. Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by *Helicobacter pylori*. *Proc Natl Acad Sci U S A* (1999) **96**:14559–64. doi:10.1073/pnas.96.25.14559
10. Lin W-C, Tsai H-F, Kuo S-H, Wu M-S, Lin C-W, Hsu P-I, et al. Translocation of *Helicobacter pylori* CagA into human B lymphocytes, the origin of mucosa-associated lymphoid tissue lymphoma. *Cancer Res* (2010) **70**:5740–8. doi:10.1158/0008-5472.CAN-09-4690
11. Fujimori K, Shimodaira S, Akamatsu T, Furihata K, Katsuyama T, Hosaka S. Effect of *Helicobacter pylori* eradication on ongoing mutation of immunoglobulin genes in gastric MALT lymphoma. *Br J Cancer* (2005) **92**:312–9. doi:10.1038/sj.bjc.6602262
12. Suzuki H, Saito Y, Hibi T. *Helicobacter pylori* and gastric mucosa-associated lymphoid tissue (MALT) lymphoma: updated review of clinical outcomes and the molecular pathogenesis. *Gut Liver* (2009) **3**:81–7. doi:10.5009/gnl.2009.3.2.81
13. Lochhead P, El-Omar E. *Helicobacter pylori* infection and gastric cancer. *Best Pract Res Clin Gastroenterol* (2007) **21**:281–97. doi:10.1016/j.bpg.2007.02.002
14. Wotherspoon AC, Ortiz Hidalgo C, Falzon MR, Isaacson PG. *Helicobacter pylori*-associated gastritis and primary B-cell gastric lymphoma. *Lancet* (1991) **338**:1175–6. doi:10.1016/0140-6736(91)92035-Z
15. Zucca E, Bertoni F, Roggero E, Bosshard G, Cazzaniga G, Pedrinis E, et al. Molecular analysis of the progression from *Helicobacter pylori*-associated chronic gastritis to mucosa-associated lymphoid-tissue lymphoma of the stomach. *N Engl J Med* (1998) **338**:804–10. doi:10.1056/NEJM199803193381205
16. Isaacson PG. Lymphomas of mucosa associated lymphoid tissue (MALT). *Am J Surg Pathol* (1992) **16**:201–5. doi:10.1097/00000478-199202000-00023
17. Cavalli F, Isaacson PG, Gascoyne RD, Zucca E. MALT lymphomas. *Hematology Am Soc Hematol Educ Program* (2001) **2001**:241–58. doi:10.1182/asheducation-2001.1.241
18. Chan JKC, Ng CS, Isaacson PG. Relationship between high-grade lymphoma and low-grade B-cell mucosa-associated lymphoid tissue lymphoma (MALToma) of the stomach. *Am J Pathol* (1990) **136**:1153–64.
19. Peng H, Du M, Diss TC, Isaacson PG, Pan L. Genetic evidence for a clonal link between low and high-grade components in gastric MALT B-cell lymphoma. *Histopathology* (1997) **30**:425–9. doi:10.1046/j.1365-2559.1997.5450786.x
20. Freeman C, Berg JW, Cutler SJ. Occurrence and prognosis of extranodal lymphomas. *Cancer* (1972) **29**:252–60. doi:10.1002/1097-0142(197201)29:1<252::AID-CNCR2820290138>3.0.CO;2-\#
21. Go JH, Kim DS, Kim TJ, Ko YH, Ra HK, Rhee JC, et al. Comparative studies of somatic and ongoing mutations in immunoglobulin heavy-chain variable region genes in diffuse large B-cell lymphomas of the stomach and the small intestine. *Arch Pathol Lab Med* (2003) **127**:1443–50. doi:10.1043/1543-2165(2003)127<1443:CSOSAO>2.0.CO;2
22. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* (2000) **403**:503–11. doi:10.1038/35000501
23. De Paepe P, De Wolf-Peeters C. Diffuse large B-cell lymphoma: a heterogeneous group of non-Hodgkin lymphomas comprising several distinct clinicopathological entities. *Leukemia* (2007) **21**:37–43. doi:10.1038/sj.leu.2404449

24. Banerjee M, Mehr R, Belelovsky A, Spencer J, Dunn-Walters DK. Age- and tissue-specific differences in human germinal center B cell selection revealed by analysis of IgVH gene hypermutation and lineage trees. *Eur J Immunol* (2002) **32**:1947–57. doi:10.1002/1521-4141(200207)32:7<1947::AID-IMMU1947>3.0.CO;2-1
25. Steiman-Shimony A, Edelman H, Barak M, Shahaf G, Dunn-Walters DK, Stott DI, et al. Immunoglobulin variable-region gene mutational lineage tree analysis: application to autoimmune diseases. *Autoimmun Rev* (2006) **5**:242–51. doi:10.1016/j.autrev.2005.07.008
26. Steiman-Shimony A, Edelman H, Hutzler A, Barak M, Zuckerman NS, Shahaf G, et al. Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: chronic activation, normal selection. *Cell Immunol* (2006) **244**:130–6. doi:10.1016/j.cellimm.2007.01.009
27. Zuckerman NS, Hazanov H, Barak M, Edelman H, Hess S, Shcolnik H, et al. Somatic hypermutation and antigen-driven selection of B cells are altered in autoimmune diseases. *J Autoimmun* (2010) **35**:325–35. doi:10.1016/j.jaut.2010.07.004
28. Zuckerman NS, Howard WA, Bismuth J, Gibson KL, Edelman H, Berrih-Aknin S, et al. Ectopic GC in the thymus of myasthenia gravis patients show characteristics of normal GC. *Eur J Immunol* (2010) **40**:1150–61. doi:10.1002/eji.200939914
29. Tabibian-Keissar H, Zuckerman NS, Barak M, Dunn-Walters DK, Steiman-Shimony A, Chowers Y, et al. B-cell clonal diversification and gut-lymph node trafficking in ulcerative colitis revealed using lineage tree analysis. *Eur J Immunol* (2008) **38**:2600–9. doi:10.1002/eji.200838333
30. Zuckerman NS, McCann KJ, Ottensmeier CH, Barak M, Shahaf G, Edelman H, et al. Ig gene diversification and selection in follicular lymphoma, diffuse large B cell lymphoma and primary central nervous system lymphoma revealed by lineage tree and mutation analyses. *Int Immunol* (2010) **22**:875–87. doi:10.1093/intimm/dxq441
31. Manske MK, Zuckerman NS, Timm MM, Maiden S, Edelman H, Shahaf G, et al. Quantitative analysis of clonal bone marrow CD19+ B cells: use of B cell lineage trees to delineate their role in the pathogenesis of light chain amyloidosis. *Clin Immunol* (2006) **120**:106–20. doi:10.1016/j.clim.2006.01.008
32. Abraham RS, Manske MK, Zuckerman NS, Sohni A, Edelman H, Shahaf G, et al. Novel analysis of clonal diversification in blood B cell and bone marrow plasma cell clones in immunoglobulin light chain amyloidosis. *J Clin Immunol* (2007) **27**:69–87. doi:10.1007/s10875-006-9056-9
33. Gurrieri C, McGuire P, Zan H, Yan X-J, Cerutti A, Albesiano E, et al. Chronic lymphocytic leukemia B cells can undergo somatic hypermutation and intraclonal immunoglobulin V(H)DJ(H) gene diversification. *J Exp Med* (2002) **196**:629–39. doi:10.1084/jem.20011693
34. Kostareli E, Sutton L, Hadzidimitriou A, Darzentas N, Kouvatsi A, Tsaftaris A, et al. Intraclonal diversification of immunoglobulin light chains in a subset of chronic lymphocytic leukemia alludes to antigen-driven clonal evolution. *Leukemia* (2010) **24**:1317–24. doi:10.1038/leu.2010.90
35. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) **23**(11):1874–84. doi:10.1101/gr.154815.113
36. Zhu D, Orchard J, Oscier DG, Wright DH, Stevenson FK. V(H) gene analysis of splenic marginal zone lymphomas reveals diversity in mutational status and initiation of somatic mutation in vivo. *Blood* (2002) **100**:2659–61. doi:10.1182/blood-2002-01-0169
37. Traverse-Glehen A, Davi F, Ben Simon E, Callet-Bauchu E, Felman P, Baseggio L, et al. Analysis of VH genes in marginal zone lymphoma reveals marked heterogeneity between splenic and nodal tumors and suggests the existence of clonal selection. *Haematologica* (2005) **90**:470–8.
38. Matolcsy A, Schattner EJ, Knowles DM, Casali P. Clonal evolution of B cells in transformation from low- to high-grade lymphoma. *Eur J Immunol* (1999) **29**:1253–64. doi:10.1002/(SICI)1521-4141(199904)29:04<1253::AID-IMMU1253>3.0.CO;2-8
39. Carlotti E, Wrench D, Matthews J, Iqbal S, Davies A, Norton A, et al. Transformation of follicular lymphoma to diffuse large B-cell lymphoma may occur by divergent evolution from a common progenitor cell or by direct evolution from the follicular lymphoma clone. *Blood* (2009) **113**:3553–7. doi:10.1182/blood-2008-08-174839
40. Green MR, Gentles AJ, Nair RV, Irish JM, Kihira S, Liu CL, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood* (2013) **121**:1604–11. doi:10.1182/blood-2012-09-457283
41. Wündisch T, Thiede C, Alpen B, Stolte M, Neubauer A. Are lymphocytic monoclonality and immunoglobulin heavy chain (IgH) rearrangement premalignant conditions in chronic gastritis? *Microsc Res Tech* (2001) **53**:414–8. doi:10.1002/jemt.1110
42. Miyamoto M, Haruma K, Hiyama T, Kamada T, Masuda H, Shimamoto F, et al. High incidence of B-cell monoclonality in follicular gastritis: a possible association between follicular gastritis and MALT lymphoma. *Virchows Arch* (2002) **440**:376–80. doi:10.1007/s00428-001-0575-8
43. Georgopoulos SD, Triantafyllou K, Fameli M, Kitsanta P, Spiliadi C, Anagnostou D, et al. Molecular analysis of B-cell clonality in *Helicobacter pylori* gastritis. *Dig Dis Sci* (2005) **50**:1616–20. doi:10.1007/s10620-005-2905-1
44. Bahler DW, Szankasi P, Kulkarni S, Tubbs RR, Cook JR, Swerdlow SH. Use of similar immunoglobulin VH gene segments by MALT lymphomas of the ocular adnexa. *Mod Pathol* (2009) **22**:833–8. doi:10.1038/modpathol.2009.42
45. Walsh SH, Rosenquist R. Immunoglobulin gene analysis of mature B-cell malignancies: reconsideration of cellular origin and potential antigen involvement in pathogenesis. *Med Oncol* (2005) **22**:327–41. doi:10.1385/MO:22:4:327
46. Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL, et al. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest* (1998) **102**:1515–25. doi:10.1172/JCI3009
47. Yamashita Y, Kajiura D, Tang L, Hasegawa Y, Kinoshita T, Nakamura S, et al. XCR1 expression and biased VH gene usage are distinct features of diffuse large B-cell lymphoma initially manifesting in the bone marrow. *Am J Clin Pathol* (2011) **135**:556–64. doi:10.1309/AJCPCTDC5PY3LXBP
48. Bende RJ, Aarts WM, Riedl RG, de Jong D, Pals ST, van Noesel CJM. Among B cell non-Hodgkin's lymphomas, MALT lymphomas express a unique antibody repertoire with frequent rheumatoid factor reactivity. *J Exp Med* (2005) **201**:1229–41. doi:10.1084/jem.20050068
49. Pimentel BJ, Stefanoff CG, Moreira AS, Seuánez HN, Zalcberg IR. Use of V H, D and J H immunoglobulin gene segments in Brazilian patients with chronic lymphocytic leukaemia (CLL). *Genet Mol Biol* (2008) **31**:643–8. doi:10.1590/S1415-47572008000400007
50. Matsuda F, Shin EK, Nagaoka H, Matsumura R, Haino M, Fukita Y, et al. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nat Genet* (1993) **3**:88–94. doi:10.1038/ng0193-88
51. Johnson TA, Rassenti LZ, Kipps TJ. Ig VH1 genes expressed in B cell chronic lymphocytic leukemia exhibit distinctive molecular features. *J Immunol* (1997) **158**:235–46.
52. Bayerl MG, Bentley G, Bellan C, Leoncini L, Ehmann WC, Palutke M. Lacunar and Reed-Sternberg – like cells in follicular lymphomas are clonally related to the centrocytic and centroblastic cells as demonstrated by laser capture microdissection. *Am J Clin Pathol* (2004) **122**:858–64. doi:10.1309/PMR86PHKK4J3RUH3
53. Hashimoto Y, Tsukamoto N, Nakahashi H, Yokohama A, Saitoh T, Handa H, et al. Hairy cell leukemia-related disorders consistently show low CD27 expression. *Pathol Oncol Res* (2009) **15**:615–21. doi:10.1007/s12253-009-9161-1
54. Nakamura-Kikuoka S, Takahi K, Tsuboi H, Toyosaki-Maeda T, Maeda-Tanimura M, Wakasa C, et al. Limited VH gene usage in B-cell clones established with nurse-like cells from patients with rheumatoid arthritis. *Rheumatology* (2006) **45**:549–57. doi:10.1093/rheumatology/kei170
55. Bahler DW, Swerdlow SH. Clonal salivary gland infiltrates associated with myoepithelial sialadenitis (Sjögren's syndrome) begin as nonmalignant antigen-selected expansions. *Blood* (1998) **91**:1864–72.
56. Brezinschek HP, Foster SJ, Brezinschek RI, Dörner T, Domiati-Saad R, Lipsky PE. Analysis of the human VH gene repertoire. Differential effects of selection and somatic hypermutation on human peripheral CD5(+)/IgM+ and CD5(-)/IgM+ B cells. *J Clin Invest* (1997) **99**:2488–501. doi:10.1172/JCI119433
57. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun A, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010) **116**:1070–8. doi:10.1182/blood-2010-03-275859
58. Perotti M, Ghidoli N, Altara R, Diotti RA, Clementi N, De Marco D, et al. Hepatitis C virus (HCV)-driven stimulation of subfamily-restricted natural IgM antibodies in mixed cryoglobulinemia. *Autoimmun Rev* (2008) **7**:468–72. doi:10.1016/j.autrev.2008.03.008

59. Mao Z, Quintanilla-Martinez L, Raffeld M, Richter M, Krugmann J, Burek C, et al. IgVH mutational status and clonality analysis of Richter's transformation: diffuse large B-cell lymphoma and Hodgkin lymphoma in association with B-cell chronic lymphocytic leukemia (B-CLL) represent 2 different pathways of disease evolution. *Am J Surg Pathol* (2007) **31**:1605–14. doi:10.1097/PAS.0b013e31804bdaf8
60. Sakuma H, Nakamura T, Uemura N, Chiba T, Sugiyama T, Asaka M, et al. Immunoglobulin VH gene analysis in gastric MALT lymphomas. *Mod Pathol* (2007) **20**:460–6. doi:10.1038/modpathol.3800758
61. Lenze D, Greiner A, Knörr C, Anagnostopoulos I, Stein H, Hummel M. Receptor revision of immunoglobulin heavy chain genes in human MALT lymphomas. *Mol Pathol* (2003) **56**:249–55. doi:10.1136/mp.56.5.249
62. Alpen B, Wündisch T, Dierlamm J, Börsch G, Stolte M, Neubauer A. Clonal relationship in multifocal non-Hodgkin's lymphoma of mucosa-associated lymphoid tissue (MALT). *Ann Hematol* (2004) **83**:124–6. doi:10.1007/s00277-003-0763-5
63. Siakantaris MP, Pangalis GA, Dimitriadou E, Kontopidou FN, Vassilakopoulos TP, Kalpadakis C, et al. Early-stage gastric MALT lymphoma: is it a truly localized disease? *Oncologist* (2009) **14**:148–54. doi:10.1634/theoncologist.2008-0178
64. De Wolf-Peeters C, Achten R. The histogenesis of large-cell gastric lymphomas. *Histopathology* (1999) **34**:71–5. doi:10.1046/j.1365-2559.1999.00602.x
65. Friedberg JW. Diffuse large B-cell lymphoma. *J Hematol Oncol* (2008) **22**:941–52. doi:10.1016/j.hoc.2008.07.002
66. Lenze D, Berg E, Volkmer-Engert R, Weiser A, Greiner A, Knörr-Wittmann C, et al. Influence of antigen on the development of MALT lymphoma. *Blood* (2006) **107**:1141–8. doi:10.1182/blood-2005-04-1722
67. Shahaf G, Barak M, Zuckerman NS, Swerdlin N, Gorfine M, Mehr R. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: a large-scale simulation study. *J Theor Biol* (2008) **255**:210–22. doi:10.1016/j.jtbi.2008.08.005
68. Salau L, Linz B, Suerbaum S, Saunders NJ, Saunders N. The diversity within an expanded and redefined repertoire of phase-variable genes in *Helicobacter pylori*. *Microbiology* (2004) **150**:817–30. doi:10.1099/mic.0.26993-0
69. Hussell T, Isaacson PG, Crabtree JE, Spencer J. The response of cells from low-grade B-cell gastric lymphomas of mucosa-associated lymphoid tissue to *Helicobacter pylori*. *Lancet* (1993) **342**:571–4. doi:10.1016/0140-6736(93)91408-E
70. Hussell T, Isaacson PG, Crabtree JE, Spencer J. *Helicobacter pylori*-specific tumour-infiltrating T cells provide contact dependent help for the growth of malignant B cells in low-grade gastric lymphoma of mucosa-associated lymphoid tissue. *J Pathol* (1996) **178**:122–7. doi:10.1002/(SICI)1096-9896(199602)178:2<122::AID-PATH486>3.0.CO;2-D
71. Ernst PB, Gold BD. The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Annu Rev Microbiol* (2000) **54**:615–40. doi:10.1146/annurev.micro.54.1.615
72. Tabibian-Keissar H, Schibby G, Michaeli M, Rakovsky-Shapira A, Azogui-Rosenthal N, Dunn-Walters DK, et al. PCR amplification and high throughput sequencing of immunoglobulin heavy chain genes from formalin-fixed paraffin-embedded human biopsies. *Exp Mol Pathol* (2012) **94**:182–7. doi:10.1016/j.yexmp.2012.08.002
73. Michaeli M, Noga H, Tabibian-Keissar H, Barshack I, Mehr R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front Immunol* (2012) **3**:386. doi:10.3389/fimmu.2012.00386
74. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* (2009) **25**:195–203. doi:10.1016/j.nbt.2008.12.009
75. Volpe JM, Cowell LG, Kepler TB. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* (2006) **22**:438–44. doi:10.1093/bioinformatics/btk004
76. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* (2007) **23**:2947–8. doi:10.1093/bioinformatics/btm404
77. Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol Lett* (2012) **148**:11–22. doi:10.1016/j.imlet.2012.08.002
78. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) **324**:807–10. doi:10.1126/science.1170020
79. Briney BS, Willis JR, McKinney BA, Crowe JE. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naïve and memory repertoires that extends across individuals. *Genes Immun* (2012) **13**:469–73. doi:10.1038/gene.2012.20
80. Jost L. Partitioning diversity into independent alpha and beta components. *Ecology* (2007) **88**:2427–39. doi:10.1890/06-1736.1
81. Chao A, Shen TJ. *Program SPADE (Species Prediction and Diversity Estimation): Program and User's Guide*. (2003). Available from: http://chao.stat.nthu.edu.tw
82. Chao A, Jost L, Chiang SC, Jiang Y-H, Chazdon RL. A two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* (2008) **64**:1178–86. doi:10.1111/j.1541-0420.2008.01010.x
83. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) **338**:67–74. doi:10.1016/j.jim.2008.06.006
84. Dunn-Walters DK, Belelovsky A, Edelman H, Banerjee M, Mehr R. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Dev Immunol* (2002) **9**:233–43. doi:10.1080/10446670310001593541
85. Dunn-Walters DK, Edelman H, Mehr R. Immune system learning and memory quantified by graphical analysis of B-lymphocyte phylogenetic trees. *Biosystems* (2004) **76**:141–55. doi:10.1016/j.biosystems.2004.05.011
86. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* (1995) **57**:289–300.
87. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) **20**:683–94. doi:10.1093/intimm/dxn026
88. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) **39**:W499–504. doi:10.1093/nar/gkr413

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 December 2013; paper pending published: 17 January 2014; accepted: 20 May 2014; published online: 03 June 2014.

*Citation: Michaeli M, Tabibian-Keissar H, Schiby G, Shahaf G, Pickman Y, Hazanov L, Rosenblatt K, Dunn-Walters DK, Barshack I and Mehr R (2014) Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas. Front. Immunol. **5**:264. doi: 10.3389/fimmu.2014.00264*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Michaeli, Tabibian-Keissar, Schiby, Shahaf, Pickman, Hazanov, Rosenblatt, Dunn-Walters, Barshack and Mehr. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

GENERAL COMMENTARY
published: 05 January 2015
doi: 10.3389/fimmu.2014.00666

Addendum: Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas

Miri Michaeli[1], Hilla Tabibian-Keissar[1,2], Ginette Schiby[2], Gitit Shahaf[1], Yishai Pickman[1], Lena Hazanov[1], Kinneret Rosenblatt[2], Deborah K. Dunn-Walters[3], Iris Barshack[2,4] and Ramit Mehr[1]*

[1] *The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel*
[2] *Department of Pathology, Sheba Medical Center, Ramat Gan, Israel*
[3] *Division of Immunology, Infection, and Inflammatory Diseases, King's College London School of Medicine, London, UK*
[4] *Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*
**Correspondence: ramit.mehr@biu.ac.il*

Edited by:
Michal Or-Guil, Humboldt University of Berlin, Germany

Reviewed by:
Jose Faro, Universidade de Vigo, Spain
Andrew M. Collins, University of New South Wales, Australia

Keywords: B-cells, Ig gene, repertoire, somatic hypermutation, diversity

A commentary on

Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas
by Michaeli M, Tabibian-Keissar H, Schiby G, Shahaf G, Pickman Y, Hazanov L, et al. Front Immunol (2014) 5:264. doi:10.3389/fimmu.2014.00264

In Section "Diversity Analysis" of this article, there were some inaccuracies in the way diversity terms were referred to. Hence, we re-wrote the section, which should read as follows.

DIVERSITY ANALYSIS

CLONES IN SAMPLES CAN BE REGARDED AS SPECIES IN HABITATS

In the case of lymphocyte clonal repertoire samples (e.g., those obtained from tissue biopsies), we treat each sample as a sample from a habitat, in which the "species" are the BCR or TCR clones found in the sample. Each of the clones may be composed of a number of different sequences. In TCR clones, all sequences are identical, but in BCR clones sequences from the same clone may be different due to somatic hypermutation, and one may choose to use only unique sequences found, or all sequences including multiplicate ones. The latter choice depends on whether identical sequences coming from different cells can be identified as such, or cannot be distinguished from sequence duplications caused by PCR amplification. If the former is true (as when using random barcoding in the PCR primers), then the number of sequences that come from different cells is known and can be used to estimate diversity. If not, then TCR diversity cannot be estimated, and BCR diversity can only be estimated based on the number of unique sequences and thus would usually only give a minimum estimate of the total diversity, as we have done in this study.

DIVERSITY INDICES

In order to quantify the diversity of clonal repertoires (such as antibody/BCR or TCR gene repertoires) in each experimental group, and later to be able to compare between two or more groups, we used diversity indices (such as the Species Richness, the Shannon entropy, or the Simpson concentration, which are indices of order 0, 1, and 2, respectively) (1). These indices take into account the number of species and (in indices of order >0) the frequency of members of a species (in our case, sequences) of each species (in our case, clone) in each habitat sample. In indices of order 0, diversity is defined simply as number of species (in our case, lymphocyte clones) in a sample. In order 1 indices, clone size (or frequency) are taken into account, as in the Shannon entropy when diversity is the sum of $[-p_i{*}\ln(p_i)]$, where i represents a species or clone and p_i represents its size (the number of members/sequences, see below). Order 2 indices attribute more weight to large clones, as in the Simpson concentration, which is the sum of p_i^2. In our studies, we used both order 1 and 2 diversity indices, i.e., the Shannon entropy and the Simpson concentration.

ESTIMATING THE FULL REPERTOIRE FROM WHICH EACH SAMPLE WAS TAKEN

Considering the large numbers of sequences observed only once in each sample, it is likely that many rare clones in an individual's original full repertoire were not detected. To account for the presence of unobserved "species" (clones), all diversity indices can be estimated for whole repertoires (rather than calculated for the sample) using the method described by Chao and Shen (2), which is based on a non-parametric estimation of diversity indices where there are undetected species. Chao and Shen's approach utilizes the concept of sample coverage to adjust the diversity indices for clones that escaped sampling. The sample coverage is estimated from the proportion of species/sequences that are observed only once within a sample.

In our Ig gene repertoire studies, the abundance data (numbers of unique sequences) of antibody clones in each sample was used to estimate the mean, standard error, and 95% confidence intervals (CI) of the diversity index of choice (order 0, 1, and 2 indices) of the full repertoire from which each sample was taken (including

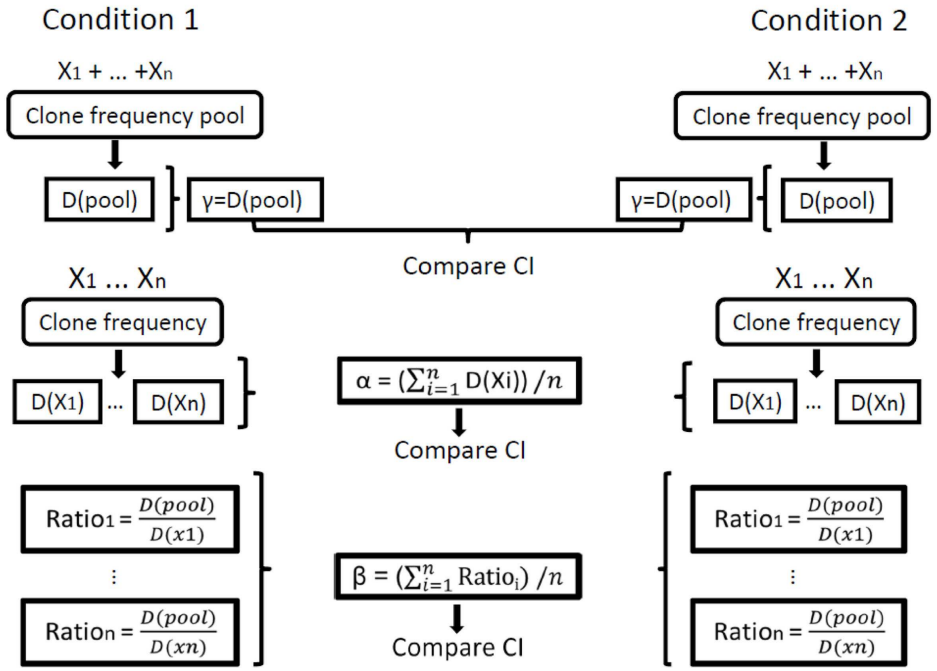


FIGURE 1 | Illustration of the calculation of repertoire diversity. First, the diversity indices (Shannon entropy, Simpson concentration, etc.) are calculated for each sample. Next, the distribution measures (α, β, γ) are calculated for each condition using the individual samples. Finally, the confidence intervals (CI) of the distribution measures are compared between the populations/conditions. Xi represents the set of clone frequencies in sample *i*, that is, the set (for all samples) of the numbers of unique sequences in all clones in sample *i* out of the total number of unique sequences in the whole sample. *D* stands for Diversity, *D*(pool) is the calculation of γ (gamma) for a sample and *D*(Xi) is the calculation of α (alpha) for sample *i*. In our calculations of order 1, we used only the Shannon entropy for D, and for order 2 we used only the Simpson concentration for D.

unobserved clones). This was done using SPADE©, a program designed for diversity calculations (2).

DIVERSITY MEASURES

From the estimated diversity indices for each sample or pool of samples, we have calculated the average alpha, beta, and gamma diversity measures for each group of samples (1). The average alpha diversity measure represents the average sample – or in our case, whole repertoire – diversity in each group of samples. In order to calculate the average alpha, we calculate the alpha diversity of each estimated repertoire using SPADE©(that is, $-\sum_{i=1}^{s}(p_i \times \ln p_i)$ for Shannon entropy, and $\sum_{i=1}^{s}\left(p_i^2\right)$ for Simpson concentration), and then average over all samples from the same group of samples. The gamma diversity measure represents the "global" repertoire diversity across all samples studied in each group of samples. It is calculated as the diversity of the pool containing all the repertoires estimated from all the samples in the same group of samples, also by SPADE©, using the same indices as for alpha.

Finally, the beta diversity measure, which represents the diversity component resulting from the variability between individual repertoires or samples, should in principle be calculated as the gamma of the group of samples divided by the average alpha of all samples in the group of samples. In order to compare between groups of samples, however, we needed to calculate CI for beta. This was done by calculating a beta measure per sample (the gamma of the group of samples divided by alpha of the sample) and then calculating the average, standard error, and CI of beta for each group of samples. Thus, comparisons can be made between the diversity measures calculated for different groups. One should keep in mind, however, that the comparisons of average alpha are based on averages of the estimates of alpha for all samples in a group, while the comparisons of gamma are based on the CI for the estimated diversity for each group. Since in all cases the 95% CI are given, if these intervals (e.g., of gamma) for two groups do not overlap, then the measures in question (e.g., gamma) of the two groups are significantly different with $p < 0.05$ under Student's t-test (**Figure 1**).

In order to allow an intuitive comparison between the diversities of each of the groups, all the diversity measures can be expressed as their number equivalents (1), which reflect the number of equally sized clones needed to produce the given value of the diversity index.

SIMILARITY ANALYSIS

In order to understand the sources for the differences in diversity between groups of samples, we used similarity analysis based on the Morisita similarity index (3). SPADE©(2) was used to calculate a similarity matrix, in which we measured each estimated individual repertoire's similarity to all other estimated individual repertoires. A value close to 1 represents high similarity between two groups, and a value close to 0 represents low similarity. The average of

similarity indices of individuals in a given group to those in another group represents the similarity index for the comparison between the two groups.

REFERENCES

1. Jost L. Partitioning diversity into independent alpha and beta components. *Ecology* (2007) **88**(10):2427–39. doi:10.1890/06-1736.1
2. Chao TJ, Shen A. Program SPADE (species prediction and diversity estimation). *Program and User's Guide* (2003). Available from: http://chao.stat.nthu.edu.tw
3. Chao A, Jost L, Chiang SC, Jiang Y-H, Chazdon RL. A two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* (2008) **64**:1178–86. doi:10.1111/j.1541-0420.2008.01010.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2014; accepted: 10 December 2014; published online: 05 January 2015.
*Citation: Michaeli M, Tabibian-Keissar H, Schiby G, Shahaf G, Pickman Y, Hazanov L, Rosenblatt K, Dunn-Walters DK, Barshack I and Mehr R (2015) Addendum: Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas. Front. Immunol. **5**:666. doi: 10.3389/fimmu.2014.00666*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2015 Michaeli, Tabibian-Keissar, Schiby, Shahaf, Pickman, Hazanov, Rosenblatt, Dunn-Walters, Barshack and Mehr. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 14 November 2013
doi: 10.3389/fimmu.2013.00374

Complementarity of binding motifs is a general property of HLA-A and HLA-B molecules and does not seem to effect HLA haplotype composition

Xiangyu Rao[1], Rob J. De Boer[1], Debbie van Baarle[2], Martin Maiers[3] and Can Kesmir[1]*

[1] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands*
[2] *Immunology Department, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, Netherlands*
[3] *National Marrow Donor Program, Minneapolis, MN, USA*

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Edward John Collins, The University of North Carolina at Chapel Hill, USA
Fernando A. Arosa, University of Beira Interior, Portugal
Dmitriy M. Chudakov, M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Russia

***Correspondence:**
Can Kesmir, Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, Netherlands
e-mail: c.kesmir@uu.nl

Different human leukocyte antigen (HLA) haplotypes (i.e., the specific combinations of HLA-A, -B, -DR alleles inherited together from one parent) are observed in different frequencies in human populations. Some haplotypes, like HLA-A1-B8, are very frequent, reaching up to 10% in the Caucasian population, while others are very rare. Numerous studies have identified associations between HLA haplotypes and diseases, and differences in haplotype frequencies can in part be explained by these associations: the stronger the association with a severe (autoimmune) disease, the lower the expected HLA haplotype frequency. The peptide repertoires of the HLA molecules composing a haplotype can also influence the frequency of a haplotype. For example, it would seem advantageous to have HLA molecules with non-overlapping binding specificities within a haplotype, as individuals expressing such an haplotype would present a diverse set of peptides from viruses and pathogenic bacteria on the cell surface. To test this hypothesis, we collect the proteome data from a set of common viruses, and estimate the total ligand repertoire of HLA class I haplotypes (HLA-A-B) using *in silico* predictions. We compare the size of these repertoires to the HLA haplotype frequencies reported in the National Marrow Donor Program (NMDP). We find that in most HLA-A and HLA-B pairs have fairly distinct binding motifs, and that the observed haplotypes do not contain HLA-A and -B molecules with more distinct binding motifs than random HLA-A and HLA-B pairs. In addition, the population frequency of a haplotype is not correlated to the distinctness of its HLA-A and HLA-B peptide binding motifs. These results suggest that there is a not a strong selection pressure on the haplotype level favoring haplotypes having HLA molecules with distinct binding motifs, which would result the largest possible presented peptide repertoires in the context of infectious diseases.

Keywords: haplotypes, HLA antigens, selection, genetic, peptide binding, bioinformatics, computational biology

INTRODUCTION

The human leukocyte antigen (HLA) genes are the most polymorphic coding loci known in humans. The HLA gene cluster is located on the major histocompatibility complex (MHC) on chromosome 6, and contains over 200 genes. The two groups of loci that contain the MHC class I and II genes dictating T cell responses are the most polymorphic. It is widely accepted that this variability is maintained by balancing selection, as individuals that are heterozygous in their HLA class I and II loci seem to have a better outcome in infections diseases [see e.g., for HIV-1 (1)]. In line with this, it has been demonstrated that in several species, and up to a certain extent in humans, females prefer to mate with males having a dissimilar HLA to increase the chance of their offspring to survive infectious diseases (2, 3). We have argued that the heterozygous advantage is on its own not enough to maintain such a large degree of polymorphism, and that the frequency dependent co-evolution with pathogens should also play a major role (4).

On the functional level the HLA polymorphism seems to be much smaller. It is well established that MHC class II molecules have largely overlapping peptide repertoires [see e.g., (5)]. In the last few years we and others showed this to be also true for class I alleles (6–9). This promiscuous peptide binding is not limited to the alleles within each locus, and extends to alleles from different MHC class I loci (7, 9). This property has important evolutionary consequences as heterozygous individuals carrying genetically different HLA molecules that nevertheless have overlapping peptide repertoires should have a diminished heterozygous advantage.

Recombinations occur frequently in the HLA region and they generate novel HLA haplotypes (10). However, the number of haplotypes found in human populations is far lower than the number of alleles observed (11), suggesting that not all the HLA haplotypes have the same chance of becoming established in human populations. Indeed, different haplotypes occur with very different frequencies in different ethnic groups/subpopulations[1,2].

[1] www.allelefrequencies.net
[2] www.nmdp.org

The "usefulness" of HLA molecules is expected to play a role in determining the haplotype frequencies: haplotypes carrying HLA molecules that are protective for certain diseases are expected to be present in relatively high frequencies in endemic areas. Indeed, several HLA haplotype disease associations have been established, e.g., in autoimmune diseases (12), squamous cell cervical cancer (13) and recurrent aphthous stomatitis (14). Along these lines, one can speculate that it should be advantageous to have MHC molecules with distinct binding specificities in a haplotype, because such haplotypes would provide an individual with more epitopes eliciting T cell responses during an infectious disease than haplotypes containing HLA molecules with similar binding motifs. This hypothesis is rather impossible to test experimentally, because one needs to determine the total set of peptides presented by a very large set of HLA molecules in cells infected by different viruses. Such experiments (e.g., eluting peptides from HLA molecules expressed by a cell) are typically done at small scales because they are time consuming. Therefore, we here study the peptide repertoires of HLA molecules that are estimated to be combined in a haplotype (i.e., the ones with a strong linkage disequilibrium) using an *in silico* approach. This approach unfortunately suffers of two main limitations. First, the quality of the peptide-HLA predictions differs between the loci. The quality of HLA-A and HLA-B peptide binding predictions is very reliable. For the other loci (HLA-C, HLA-DR, etc.) the predictions are of low quality (15), and are therefore left out of this analysis. Second, having distinct HLA binding motifs might also be important in the context of cancer or autoimmunity, however, it is very complicated to determine the role of HLA haplotypes in these diseases. Therefore, we perform our analysis explicitly for infectious diseases, and focus on the peptides that can be presented from common viruses only. Still, we think that our study is a first step to investigate the role of HLA binding motifs in the evolution of HLA haplotype frequencies.

MATERIALS AND METHODS

NMDP DATA

All estimated HLA-A-B haplotype frequencies were downloaded from the National Marrow Donor Program (NMDP) database, which was established to develop and maintain a registry of HLA-typed volunteer unrelated donors for patients requiring a hematopoietic stem cell transplant (16)[3]. Four predominant US ethnic and racial groups were included in this data set: European Americans, African Americans, Asians, and Hispanic (17). Haplotype frequencies were estimated separately for each ethnic group using an implementation of the expectation maximization (EM) algorithm (18, 19).

The linkage disequilibrium, D, between two alleles in each haplotype was expressed as the difference between the observed and expected haplotype frequency: $D = f_{AB} - f_A f_B$, where f_{AB} is the observed (estimated) haplotype frequency, f_A is the allele frequency of the HLA-A molecule in the haplotype and f_B is the allele frequency of the HLA-B molecule in this haplotype. D is easy to calculate, but has the disadvantage of depending on the frequency of the alleles. In order to overcome this drawback, the normalized measure, D′, was calculated a $D' = \frac{D}{D_{\max}}$, where $D_{\max}$ is the lesser of f_{AB} or $(1 - f_A)(1 - f_B)$ when $D < 0$ and is the lesser of $f_A(1 - f_B)$ or $f_B(1 - f_A)$ when $D > 0$. The advantage of this measure of disequilibrium is that it ranges between −1 and 1, regardless of the allelic frequencies in the sample. $|D'| = 1$ indicates complete LD and $D' = 0$ corresponds to total absence of LD.

Linkage disequilibrium statistics were calculated for each haplotype to identify the haplotypes that have a significantly positive D'.

HLA LIGAND PREDICTION

To be able to perform our analysis for as many as possible HLA molecules, we used NetMHCpan (15) to predict peptide-HLA binding affinity. NetMHCpan assigns to each peptide-HLA pair a predicted IC50 value, indicative of the predicted binding affinity. To assess whether a peptide binds to an HLA molecule depends on the choice of binding threshold, and the optimal threshold has been discussed (20). If one assumes that all HLA molecules use a fixed threshold, one can use the default threshold of 500 nM (21, 22), otherwise a 5000 nM threshold can be used to allow for the comparison of more weakly binding peptides. However, using a fixed threshold to define predicted binders result in large differences in the predicted repertoire sizes between HLA molecules. For instance using a fixed threshold of 500 nM, the HLA repertoire sizes range between 20 and 6574 peptides for the viral set listed in Table S2 in Supplementary Material. As such a variance could introduce large biases in our analysis, we defined the 1% top-ranking peptides as candidate binders for each HLA molecule. This gives each HLA molecule the same ligand repertoire size (i.e., 570 binders per HLA molecule for the viral set listed in Table S2 in Supplementary Material). To check the consistency of our results with respect to these parameters, we repeated every analysis with the fixed threshold of 500 nM. All results presented below were derived using the threshold of 1% to define candidate binders, and remain similar for a fixed threshold, unless mentioned otherwise. To test whether our results change if one were to use a much larger data set, we also generated predicted binders using a much larger set of viruses (see below), in which case each HLA molecule had 60-fold more binders.

VIRAL DATA

The proteomes of 17 common human viruses were downloaded from the European Bioinformatics Institute website[4] (downloads were made in October 2006, listed in Table S2 in Supplementary Material) as the source of potential HLA ligands. To extend this data set, we downloaded another set of proteomes (downloads were made in October 2008)[4] from viruses that are known to infect mammalians ($n = 904$). We used the HLA-peptide binding predictors (see above) to screen all possible unique virus-derived 9-mer peptides.

PEPTIDE REPERTOIRE OVERLAP

We define the peptide repertoire overlap between two HLA alleles in the same HLA-A-B haplotype, F_p, as the fraction of overlapping ligands between these two HLA class I molecules among

[3] http://bioinformatics.nmdp.org

[4] www.ebi.ac.uk

all of their ligands: $F_p = \frac{A \bigcap B}{A \bigcup B}$, where A and B are the ligand sets for HLA-A and -B molecule, respectively. $F_p = 1$ implies that the HLA-A and HLA-B molecules belonging to the same haplotype have the same epitope repertories while $F_p = 0$ indicates completely different peptide repertoires for two HLA molecules.

RESULTS AND DISCUSSION

DETERMINING HLA-A-B HAPLOTYPES

The "true" HLA haplotypes can be determined either by molecular haplotyping or family-based segregation studies (23–25). However, both approaches are expensive and laborious, and therefore, statistical methods are typically used to infer haplotypes from datasets covering large population of individuals with known HLA genotypes (26). Several methods have been proposed to infer HLA haplotypes from genotype data, and in recent studies the performance of two most commonly used approaches, EM algorithm based (implemented in Arlequin V3.0), and the Bayesian algorithm based (implemented in PHASE V2.1.1), have been compared (27, 28). Unfortunately, neither of the methods could infer all of the known haplotypes: incorrect haplotypes were estimated in more than 30% of the cases. However, once the sample size increases, the power of these statistical methods is expected to increase tremendously.

National marrow donor program[2] provides, to our knowledge, the largest repository of HLA-typed donors. Here use of statistical methods should become more reliable (16): for the HLA-A-B haplotype, the total chromosome counts (2N) for the four major ethnic groups exceeds 2000. On the NMDP webpage[3], the high-resolution allele and haplotype frequencies [estimated by EM method, (18, 19)] are available (17). Focusing on HLA-A-B haplotypes, the most common haplotypes found in US population (separated into four main ethnic groups) are summarized in **Table 1** (adopted from bioinformatics.nmdp.org, December 2007 version). Alternatively Allele frequencies web server[1], provides allele frequencies established in smaller, but probably better defined studies (29).

In the NMDP database 660 possible haplotypes are reported for European Americans. However, many of these haplotypes are bound to be falsely predicted, e.g., due to the limited number of individuals carrying particular combinations of specific HLA molecules. To decrease the amount of wrongly identified haplotypes in our analysis, we apply a rather strict criterion for considering a predicted haplotype as a "true" haplotype: we demand a positive LD value that is significantly different than zero ($p < 0.01$, see Materials and Methods). This criterion decreases the number of haplotypes to 60 for European Americans. These 60 haplotypes are estimated to cover 58% of the population (see **Table 2**), i.e., current statistical methods and data sets (even the large repositories like NMDP) remain rather limited in providing the HLA haplotype diversity of a population. For other ethnical groups, the number of reliable haplotypes drops to 30–40 per ethnic group, even though the number of possible haplotypes was the same or higher (see **Table 2** and footnote text 3). The population coverage in non-European groups was lower, 30–40% of their respective populations, possibly due to the lower number of individuals with known HLA-typing. All together we detected 120 reliable unique haplotypes by summing over these ethnical groups (the full list of haplotypes can be found in Table S1 in Supplementary Material).

Table 1 | Occurrences of the three most common HLA-A-B frequency ranked haplotypes in four major ethnic groups in US (adopted from bioinformatics.nmdp.org).

		EUR		AFA		API		HIS	
HLA-A	**HLA-B**	***F* (%)**	**Rank**	***F* (%)**	**Rank**	***F* (%)**	**Rank**	***F* (%)**	**Rank**
0101	0801	9.55	1	1.50	6	0.41	46	2.21	2
0201	4402	5.70	3	1.33	9	0.17	130	1.94	4
0201	4501	0.05	200	1.66	3	–	–	0.23	105
0201	5101	2.00	9	0.61	26	0.91	25	2.20	3
0207	4601	–	–	–	–	3.34	2	–	–
0301	0702	6.01	2	1.73	2	0.26	82	1.92	5
2902	4403	2.38	7	1.08	15	0.03	433	2.54	1
3001	4201	–	–	2.96	1	–	–	0.40	50
3303	4403	–	–	0.09	261	2.94	3	0.16	156
3303	5801	0.08	162	0.28	88	4.53	1	0.10	230

EUR, Caucasian; AFA, African; API, Asian; HIS, Hispanic. F stands for population frequency in percentages.

Table 2 | Numbers of different haplotypes with a significantly positive LD in four major US ethnic groups.

Ethnicity	**Haplotype # (%)**
EUR	60 (57.7)
API	34 (35.4)
HIS	43 (33.5)
AFA	43 (33.7)

Population coverage in percentages is given within parenthesis.

PEPTIDE REPERTOIRE OF AN HAPLOTYPE

Having identified the HLA-A-B haplotypes for the US population, we next estimated the overlaps between peptide repertoires of HLA-A and -B molecules that belong to the same haplotype.

We used an *in silico* approach and predicted the peptide repertoire of all HLA-A and HLA-B alleles that are part of the 120 predicted haplotypes, using the proteomes of common viruses (see Table S2 in Supplementary Material) and HLA-peptide binding predictor NetMHCpan (15, 30) (see Materials and Methods). NetMHCpan is the only prediction system available right now that can reliably predict the peptide binding affinities for the large set of HLA-A and HLA-B molecules we are taking into account in this study. The analysis of the 120 significant haplotypes demands predictions for 39 HLA-A and 63 HLA-B molecules. This predictor assigns an IC50 value to each peptide-HLA pair, which can be used as a predicted binding affinity. Using the widely accepted IC50 value of 500 nM as a threshold to distinguish binders from non-binders, generated a large variation in the predicted repertoire sizes of different HLA molecules (20–6574 peptides for the viral set listed in Table S2 in Supplementary Material), which could strongly bias our results. As the physiologically relevant IC50 values are difficult to estimate for each HLA molecule, we have chosen

a simplified approach and define the peptide repertoire as the top 1% peptides with the highest HLA binding affinities for each HLA molecule. This approach removes the potential bias introduced by different repertoire sizes, but ignores the fact that some HLA molecules can be more specific than others, and therefore present much fewer peptides.

The unique haplotypes listed in Table S1 in Supplementary Material have an average peptide overlap of 1.1% (with a standard deviation of 1.8%, and a median of 0.44%) using top 1% threshold. The distribution of the overlaps is given in **Figure 1A**, and varies somewhat among the haplotypes. Out of 120 haplotypes, 36 (30%) have non-overlapping peptide repertoires, at least for the common viruses that we tested (see Table S1 in Supplementary Material). Only for five haplotypes is the repertoire overlap higher than 5%, and HLA-A0101-B1517 with an overlap of 11.8% is the highest. HLA-A0101-B1517 is a rare haplotype and occurs only in Asian Americans with a population frequency of 0.4%. To test whether or not rare haplotypes tend to have larger overlaps than common haplotypes, we weighted the overlaps found in **Figure 1A** with the population frequency of the HLA-A-B haplotypes (**Figure 1B**). Since the weighted overlaps remain very similar to the unweighted overlaps (**Figures 1A,B**), there is no evidence for a trend of rare haplotypes having the largest overlaps. In line with this, the frequency of a haplotype is only weakly correlated with the degree of peptide overlap between the haplotype's HLA-A and HLA-B molecules ($r = -0.08$, $p = 0.4$, Spearman rank correlation). Apparently, there is no selection pressure increasing the frequency of the haplotypes with a small peptide repertoire overlap.

These results were obtained using the top 1% peptides with the highest HLA binding affinities as the set of presented peptides per molecule. Using the set of common viruses listed in Table S2 in Supplementary Material, this threshold results in approximately 570 predicted binders per HLA molecule. To test whether the results presented in **Figure 1A** would be sensitive to the number of peptides, we collected proteomes for a much larger set of mammalian viruses ($n = 904$), which contains approximately three million unique peptides of nine amino acids. Using the same 1% threshold for this larger set, we predict for each HLA molecule the extended peptide repertoire and calculate the overlaps as explained before. The distribution of the overlaps hardly changes despite the fact that we enlarged the presented peptide repertoire 60-fold per molecule (see Figure S1 in Supplementary Material). As the results presented in **Figure 1A** seem fairly insensitive to the number of peptides used, we perform the rest of the analysis on the small data set of common viruses.

To test whether or not HLA binding motifs affect haplotype compositions requires comparison of the peptide overlaps of "true" haplotypes with those of "random" haplotypes. To do this, we reshuffled HLA-A and HLA-B molecules in the 120 "true" haplotypes to calculate an expected peptide repertoire overlap for randomly made haplotypes. Although the HLA molecules in random haplotypes can have overlaps up to 28% in their peptide repertoires (see **Figure 1C** and results not shown), the distribution of the overlaps is not significantly different from the distribution given in **Figure 1A**. The set of random haplotypes was generated 100 times, and in none of these cases was the distribution significantly different from the one given in **Figure 1A** (using a Kolmogorov–Smirnov test). Finally, we calculated a weighted peptide repertoire overlap for the random haplotypes by assuming that the frequency of a random haplotype is simply the multiplication of the frequency of their HLA-A and HLA-B alleles (i.e., assuming a complete lack of linkage disequilibrium). Again, the distribution of weighted overlap of random haplotypes (**Figure 1D**) is not different from that of the real haplotypes (**Figure 1B**). Taken together, these results suggest that HLA-A and -B in pairs in general have distinct peptide binding preferences, and that a small overlap is not a unique property of the HLA molecules having a strong linkage disequilibrium.

The overlap distribution presented in **Figure 1** seems to be in contradiction with earlier results, which estimated cross loci peptide overlaps of 23–44% (7, 9). However, this overlap was estimated at the population level, i.e., these percentages reflect the fraction of the peptide repertoire of an HLA-A molecule which is also expected to be presented by at least one HLA-B molecule in the population (and vice versa). Within an individual having maximally two different HLA-A and HLA-B molecules, the overlaps should remain much lower than the population based overlaps. In addition, one needs to realize that the low overlaps presented in **Figure 1** depend on the threshold used to define the peptide repertoire of an HLA molecule. When we use a higher threshold of 2 or 5% the average overlap increases to 2.3 and 5.9%, respectively (with the standard threshold of 1%, the average overlap was 1.1%, see above). However, the choice of the threshold hardly affects our main result, namely that the HLA-A and -B pairs that are in a linkage disequilibrium do not have more distinct binding motifs than random HLA-A/B pairs (**Figure 1**).

CONCLUSION

We hypothesized that it should be advantageous to have HLA molecules with distinct binding specificities combined in a haplotype, because during a viral infection such haplotypes would give an individual a larger epitope repertoire than haplotypes containing HLA molecules with similar binding motifs during a viral infection. To test this hypothesis, we used the high-resolution data available in the NMDP database on haplotype frequencies, and employed state of the art peptide-HLA binding prediction tools. We find that for all the haplotypes we could reliably identify in the US population, their HLA-A and HLA-B molecules present largely distinct set of peptides (**Figure 1A**). However, this turned out to be a generic property of HLA-A and HLA-B molecules: when we compared random HLA-A and -B pairs we find a very similar distribution of the presented peptide overlaps (**Figure 1C**). Moreover, there is no evidence for selection as there is no correlation between the population frequency of the HLA-A-B haplotypes and the overlap in the peptide repertoires of their HLA-A and HLA-B molecules. Taken together, these results suggest the complementarity of binding motifs is a general property of HLA-A and HLA-B molecules, and that complementarity does not affect the HLA haplotype composition. We were not able to specifically test the effect of complimentary binding motifs in the context of autoimmunity and cancer, as for both cases it remains unclear which set of human proteins should be taken as possible auto antigens. Complimentary binding motifs are expected to increase the

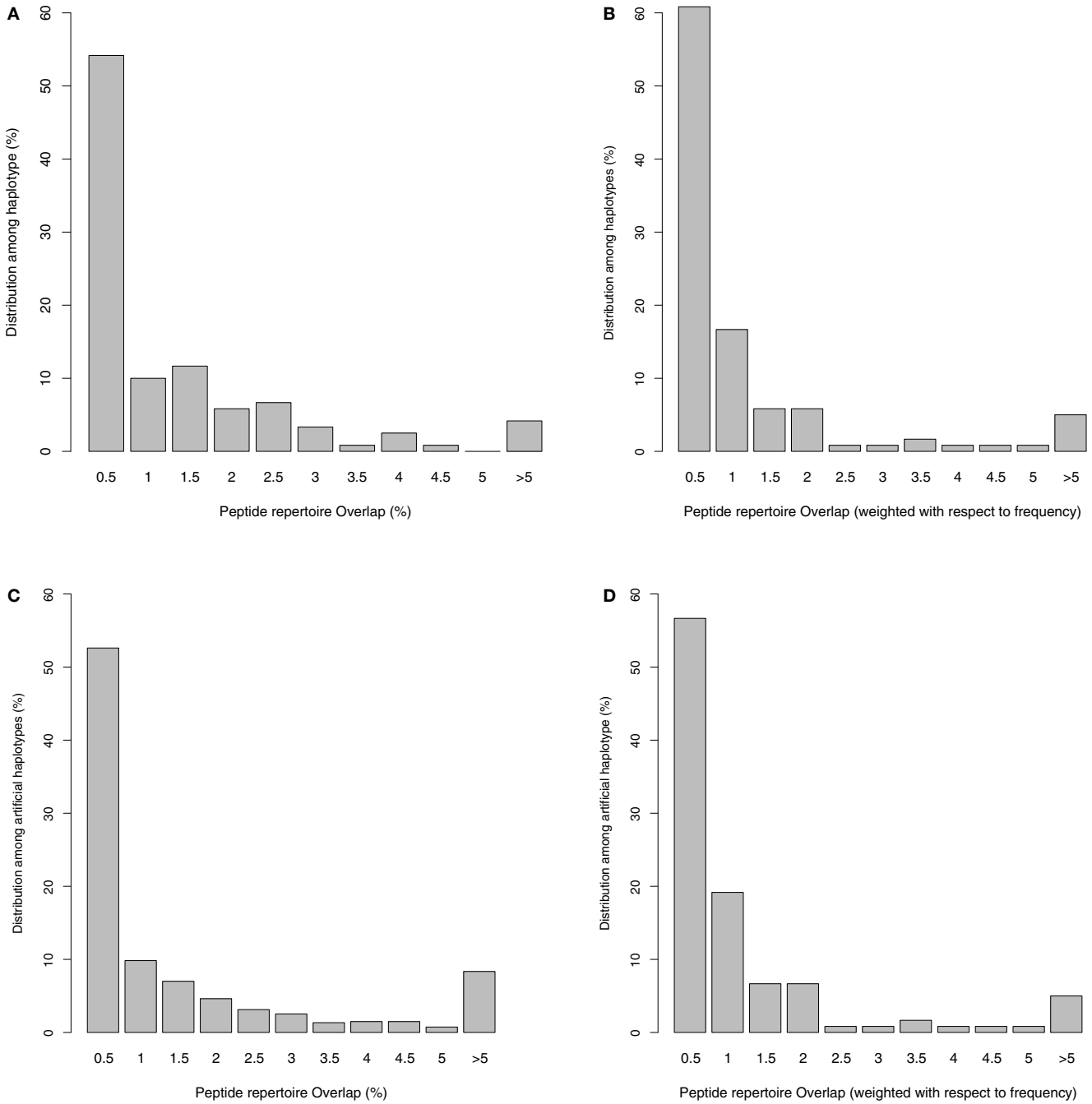


FIGURE 1 | Distribution of the presented peptide overlap (given in percentages, see Materials and Methods) between HLA-A and HLA-B molecules belonging to the same haplotype. **(A)** The distribution of the peptide overlaps for the 120 unique haplotypes we identified in the US population. **(B)** The distribution in A is weighted with the frequency of the haplotype in the population. If a haplotype was found in more than one ethnicity in US, we have taken the maximum frequency into account, normalizing the frequencies such that the sum remains one. In **(C)** we plot the overlaps found in artificially generated haplotypes (created by reshuffling the HLA-A and -B molecules from the 120 unique haplotypes). This plot is representative of 100 sets of artificially generated haplotypes. **(D)** The weighted peptide overlaps for the artificial haplotypes, where we estimate the frequency of the haplotype as the multiplication of the frequency of HLA-A and HLA-B molecule in the population (normalized to let the sum of the frequencies of the artificial haplotypes remain one).

number of potential self antigens, which could increase the risk of autoimmunity. Finally, the frequency of an HLA haplotype is determined by complex interactions with many different factors, one example is the correlation between birth weight and particular haplotypes (31). Our results suggest that complimentary binding motifs of HLA molecules during viral infections play a minor role,

if any, compared to these other factors in the evolution of HLA haplotype frequencies.

ACKNOWLEDGMENTS

This study was financially supported by the University of Utrecht through a High Potential grant. The funders had no role in study design, data collection, and analysis, decision to publish or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2013.00374/abstract

REFERENCES

1. Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* (1999) **283**:1748–52. doi:10.1126/science.283.5408.1748
2. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK. The nature of selection on the major histocompatibility complex. *Crit Rev Immunol* (1997) **17**:179–224. doi:10.1615/CritRevImmunol.v17.i2.40
3. Boehm T, Zufall F. MHC peptides and the sensory evaluation of genotype. *Trends Neurosci* (2006) **29**:100–7. doi:10.1016/j.tins.2005.11.006
4. De Boer RJ, Borghans JAM, van Boven M, Kesmir C, Weissing FJ. Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. *Immunogenetics* (2004) **55**:725–31. doi:10.1007/s00251-003-0629-y
5. O'Sullivan D, Arrhenius T, Sidney J, Del Guercio MF, Albertson M, Wall M, et al. On the interaction of promiscuous antigenic peptides with different DR alleles. Identification of common structural motifs. *J Immunol* (1991) **1950**(147):2663–9.
6. Axelsson-Robertson R, Weichold F, Sizemore D, Wulf M, Skeiky YA, Sadoff J, et al. Extensive major histocompatibility complex class I binding promiscuity for *Mycobacterium tuberculosis* TB10.4 peptides and immune dominance of human leucocyte antigen (HLA)-B*0702 and HLA-B*0801 alleles in TB10.4 CD8 T-cell responses. *Immunology* (2010) **129**:496–505. doi:10.1111/j.1365-2567.2009.03201.x
7. Frahm N, Yusim K, Suscovich TJ, Adams S, Sidney J, Hraber P, et al. Extensive HLA class I allele promiscuity among viral CTL epitopes. *Eur J Immunol* (2007) **37**:2419–33. doi:10.1002/eji.200737365
8. Nakagawa M, Kim KH, Gillam TM, Moscicki A-B. HLA class I binding promiscuity of the CD8 T-cell epitopes of human papillomavirus type 16 E6 protein. *J Virol* (2007) **81**:1412–23. doi:10.1128/JVI.01768-06
9. Rao X, Hoof I, Costa AI, van Baarle D, Kesmir C. HLA class I allele promiscuity revisited. *Immunogenetics* (2011) **63**:691–701. doi:10.1007/s00251-011-0552-6
10. Carrington M. Recombination within the human MHC. *Immunol Rev* (1999) **167**:245–56. doi:10.1111/j.1600-065X.1999.tb01397.x
11. Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, et al. Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *J Immunol* (1992) **1950**(148):249–58.
12. Smith WP, Vu Q, Li SS, Hansen JA, Zhao LP, Geraghty DE. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. *Genomics* (2006) **87**:561–71. doi:10.1016/j.ygeno.2005.11.020
13. Madeleine MM, Johnson LG, Smith AG, Hansen JA, Nisperos BB, Li S, et al. Comprehensive analysis of HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 loci and squamous cell cervical cancer risk. *Cancer Res* (2008) **68**:3532–9. doi:10.1158/0008-5472.CAN-07-6471
14. Albanidou-Farmaki E, Deligiannidis A, Markopoulos AK, Katsares V, Farmakis K, Parapanissiou E. HLA haplotypes in recurrent aphthous stomatitis: a mode of inheritance? *Int J Immunogenet* (2008) **35**:427–32. doi:10.1111/j.1744-313X.2008.00801.x
15. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* (2009) **61**:1–13. doi:10.1007/s00251-008-0341-z
16. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* (2007) **68**:779–88. doi:10.1016/j.humimm.2007.04.005
17. Kollman C, Maiers M, Gragert L, Müller C, Setterholm M, Oudshoorn M, et al. Estimation of HLA-A, -B, -DRB1 haplotype frequencies using mixed resolution data from a National Registry with selective retyping of volunteers. *Hum Immunol* (2007) **68**:950–8. doi:10.1016/j.humimm.2007.10.009
18. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* (1995) **56**:799–810.
19. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* (1995) **12**:921–7.
20. MacNamara A, Kadolsky U, Bangham CR, Asquith B. T-cell epitope prediction: rescaling can mask biological variation between MHC molecules. *PLoS Comput Biol* (2009) **5**:e1000327. doi:10.1371/journal.pcbi.1000327
21. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, et al. Sensitive quantitative predictions of peptide-MHC binding by a "Query by Committee" artificial neural network approach. *Tissue Antigens* (2003) **62**:378–84. doi:10.1034/j.1399-0039.2003.00112.x
22. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* (2003) **12**:1007–17. doi:10.1110/ps.0239403
23. Crawford DC, Nickerson DA. Definition and clinical importance of haplotypes. *Annu Rev Med* (2005) **56**:303–20. doi:10.1146/annurev.med.56.082103.104540
24. Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, et al. Conversion of diploidy to haploidy. *Nature* (2000) **403**:723–4. doi:10.1038/35002251
25. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet* (2001) **28**:361–4. doi:10.1038/ng582
26. Niu T. Algorithms for inferring haplotypes. *Genet Epidemiol* (2004) **27**:334–47. doi:10.1002/gepi.20024
27. Bettencourt BF, Santos MR, Fialho RN, Couto AR, Peixoto MJ, Pinheiro JP, et al. Evaluation of two methods for computational HLA haplotypes inference using a real dataset. *BMC Bioinformatics* (2008) **9**:68. doi:10.1186/1471-2105-9-68
28. Castelli EC, Mendes-Junior CT, Veiga-Castelli LC, Pereira NF, Petzl-Erler ML, Donadi EA. Evaluation of computational methods for the reconstruction of HLA haplotypes. *Tissue Antigens* (2010) **76**:459–66. doi:10.1111/j.1399-0039.2010.01539.x
29. Middleton D, Menchaca L, Rood H, Komerofsky R. New allele frequency database: http://www.allelefrequencies.net. *Tissue Antigens* (2003) **61**:403–7. doi:10.1034/j.1399-0039.2003.00062.x
30. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* (2007) **2**:e796. doi:10.1371/journal.pone.0000796
31. Capittini C, Pasi A, Bergamaschi P, Tinelli C, De Silvestri A, Mercati MP, et al. HLA haplotypes and birth weight variation: is your future going to be light or heavy? *Tissue Antigens* (2009) **74**:156–63. doi:10.1111/j.1399-0039.2009.01282.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 July 2013; paper pending published: 28 August 2013; accepted: 31 October 2013; published online: 14 November 2013.
*Citation: Rao X, De Boer RJ, van Baarle D, Maiers M and Kesmir C (2013) Complementarity of binding motifs is a general property of HLA-A and HLA-B molecules and does not seem to effect HLA haplotype composition. Front. Immunol. **4**:374. doi: 10.3389/fimmu.2013.00374*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Rao, De Boer, van Baarle, Maiers and Kesmir. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 30 April 2014
doi: 10.3389/fimmu.2014.00132

Receptor pre-clustering and T cell responses: insights into molecular mechanisms

Mario Castro[1]*, Hisse M. van Santen[2]*, María Férez[2], Balbino Alarcón[2], Grant Lythe[3] and Carmen Molina-París[3]

[1] Grupo de Dinámica No-Lineal and Grupo Interdisciplinar de Sistemas Complejos (GISC), Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas, Madrid, Spain
[2] Departamento de Biología Celular e Inmunología, Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas, Universidad Autónoma de Madrid, Madrid, Spain
[3] Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds, UK

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Daniel Coombs, University of British Columbia, Canada
Barbara Szomolay, University of Warwick, UK

***Correspondence:**
Mario Castro, Universidad Pontificia Comillas, C/ Alberto Aguilera 25, Madrid E28015, Spain
e-mail: marioc@upcomillas.es;
Hisse M. van Santen, Centro Biología Molecular Severo Ochoa, Calle Nicolás Cabrera 1, Campus de Cantoblanco, Madrid 28049, Spain
e-mail: hvansanten@cbm.csic.es

T cell activation, initiated by T cell receptor (TCR) mediated recognition of pathogen-derived peptides presented by major histocompatibility complex class I or II molecules (pMHC), shows exquisite specificity and sensitivity, even though the TCR–pMHC binding interaction is of low affinity. Recent experimental work suggests that TCR pre-clustering may be a mechanism via which T cells can achieve such high sensitivity. The unresolved stoichiometry of the TCR makes TCR–pMHC binding and TCR triggering, an open question. We formulate a mathematical model to characterize the pre-clustering of T cell receptors (TCRs) on the surface of T cells, motivated by the experimentally observed distribution of TCR clusters on the surface of naive and memory T cells. We extend a recently introduced stochastic criterion to compute the timescales of T cell responses, assuming that ligand-induced cross-linked TCR is the minimum signaling unit. We derive an approximate formula for the mean time to signal initiation. Our results show that pre-clustering reduces the mean activation time. However, additional mechanisms favoring the existence of clusters are required to explain the difference between naive and memory T cell responses. We discuss the biological implications of our results, and both the compatibility and complementarity of our approach with other existing mathematical models.

Keywords: T cell receptor, clustering, stochastic dynamics, signaling, naive T cells, memory T cells

1. INTRODUCTION

A hallmark of the adaptive immune system is the ability of T cells, making use of the T cell receptors (TCRs) on their surface, to recognize a given agonist peptide–MHC ligand complex (pMHC) with high sensitivity (1). Some aspects of TCR–pMHC molecular interactions that are of current research interest are the frequency of encounters between T cells and the agonist pMHC, how cell–cell interactions determine the activation of lymphocytes (2), how early interactions change the state of the T cell receptor (3), what are the mechanisms of modulation of receptor–ligand interactions at cell–cell interfaces (4), and how protein organization in the cell membrane (for instance, protein islands or lipid rafts) affects the recognition process (5). Some recent experiments have explored the role of dimensionality on T cell activation and have highlighted the significance of the events taking place at the receptor level [see Refs. (1) and (6) for comprehensive reviews].

These open questions have been addressed with the use of mathematical modeling. Different theories can be classified according to the level of description (7). At the individual TCR–pMHC bond level, the kinetic proof-reading model (8) assumes that the TCR needs to undergo a series of consecutive (phosphorylation) steps before being triggered. Also at the TCR level, the optimal dwell time model (9) reconciles the concurrence of different timescales, providing an *optimal* timescale between the very short times related to the off rate of TCR–pMHC binding, and the long times related to kinetic proof-reading mechanisms. The TCR occupancy model (10) considers the cell as a *counting device* in which multiple TCR–pMHC interactions are required to activate a T cell. In a similar fashion, the serial triggering model (11) proposed that the same pMHC can engage *serially* different TCRs. This model enriches the viewpoint of the TCR occupancy model, by giving greater relevance to the role of the pMHC itself. Finally, the serial encounter model (12) and the confinement time model (13) combine several of the ideas above and provide some appealing explanations by relaxing some restrictions in those models.

While antigen presenting cells (APCs), such as dendritic cells or B cells, present 10^3–10^4 times more self-pMHC than antigenic pMHC, self-pMHC ligands by themselves do not usually elicit a T cell response, even though their affinity for TCR$\alpha\beta$ is only 10 times lower than the affinity of the antigenic pMHC (14). This illustrates how a small difference in affinity results in high specificity, when there is only a few antigenic pMHC molecules in a background of self-pMHC ligands (15).

The T cell signaling process begins with (extracellular) TCR–pMHC binding, followed by phosphorylation of the intracellular ITAM domains of the TCR–CD3 complex. When a TCR binds a pMHC molecule, the TCR$\alpha\beta$ hetero-dimer binds the peptide,

while the CD4 or CD8 co-receptor binds the MHC molecule. The binding of the co-receptor activates the tyrosine kinase Lck, which phosphorylates the ITAMs of the CD3 complex. ITAM phosphorylation allows recruitment of intracellular signaling components that mediate downstream signaling events (16).

It has recently been suggested that, contrary to what happens in TCR micro-clusters and the immunological synapse, clustering is not only induced by the ligand but by an *avidity maturation* mechanism (or pre-clustering) (17), allowing the aggregation of chains of TCRs as long as 20 units (around 200 nm long), and referred to as *nano-clusters* (3, 18). Specifically, multimeric TCR–CD3 complexes are activated at low agonistic pMHC concentrations and monomeric TCRs remain unaffected at low ligand concentration. The TCR nano-clusters could enhance T cell sensitivity by the mechanisms proposed in the models of T cell activation (7), as their existence would reduce the time needed for two (or more) receptors to aggregate (by diffusion). This pre-cluster formation could be explained by three different mechanisms (3):

- Multimeric complexes (or clusters) enhance the TCR *avidity* toward the ligand, which is expressed in clusters on the surface of APCs (19–21). At low ligand concentration, only multimeric TCR clusters are bound to ligand, as TCR monomers require higher ligand concentration. Monomeric TCRs might only be activated at high agonist doses.
- Multimeric complexes allow the propagation of the activation signal from ligand-bound TCR$\alpha\beta$ to neighboring receptors in the same TCR *multimer*.
- Linear arrays of multimeric TCR complexes help a single pMHC serially trigger several receptors (11).

The existence of these nano-clusters does not exclude additional mechanisms of T cell activation, as long as they involve the *cooperation* of receptors when they aggregate. Thus, while models such as kinetic proof-reading [and improvements as described in Ref. (22)] operate at the level of a single receptor, other models might be used in combination with the fact that the pre-cluster distribution of naive and memory T cells is different.

Additionally, the fact that the TCR stoichiometry has not been resolved under physiological conditions, yet, makes it even more difficult to understand, at a molecular level, the dynamics of TCR pre-clustering (23). TCR pre-clustering could be an example of a more general mechanism of membrane-bound molecular pre-clustering, as clustering prior to cell–cell interaction has also been observed on the surface of APCs (19–21). It is worth mentioning that monomeric TCRs can still be activated at increasing ligand concentrations, thus, conferring the T cell with a capacity to generate a dose-dependent response at very high pMHC doses, when multimeric TCR–CD3 complexes are already saturated (18). Such mechanisms have been previously described for chemotactic bacteria, as a cellular mechanism to control sensitivity (24).

Various mechanisms have already been suggested, at the population, cellular or molecular level, to explain the capacity of T cells to respond, faster and more strongly, to a second antigenic encounter. However, the underlying mechanisms of the observed changes in the sensitivity of the T cell for pMHC ligand-mediated TCR stimulation (25) have not yet been clearly elucidated. Interestingly, the distribution of clusters in naive and memory T cells is different: memory T cells accommodate larger linear TCR clusters than naive ones. This could explain why memory T cells elicit more rapid responses than naive T cells (17) (see **Figure 1** below).

In this paper, we explore the consequences of TCR pre-clustering in signaling and in distinguishing naive from memory T cell responses. We present some experimentally obtained distributions of TCR clusters for both types of cells (see **Figure 1**), and two complementary theoretical models: (i) a simple model of receptor oligomerization that describes cluster size distributions, and (ii) a generalization of the stochastic T cell response criterion of Ref. (26), to accommodate the hypothesis that the minimum signaling unit is composed of a TCR receptor cluster that is bound by the same cross-linked multivalent ligand. We find that this signaling unit is able to discriminate between agonist and antagonist pMHC ligands (with greater sensitivity than in the monomeric case), and to explain some of the advantages that higher cluster sizes can provide to memory T cells. The model also points at the need to invoke additional cooperativity mechanisms, to explain the experimentally observed role of clustering in T cell responses (27). Finally, this model of ligand-induced TCR cross-linking can be relevant in physiological conditions, according to the defective ribosomal products (DRiP) hypothesis (28, 29), which provides a rapid source of peptide precursors to optimize immuno-surveillance of pathogens and tumors (30).

2. MATHEMATICAL MODELING OF TCR PRE-CLUSTERING AND T CELL ACTIVATION

2.1. MODEL 1: T CELL RECEPTOR PRE-CLUSTERING

The TCR–CD3 complex consists of the pMHC binding TCR$\alpha\beta$ hetero-dimer, associated with the hetero-dimers CD3$\gamma\epsilon$ and CD3$\delta\epsilon$, and the homo-dimer CD3$\zeta\zeta$. Binding of a stimulating pMHC ligand by the extracellular domain of TCR$\alpha\beta$ results in conformational changes in the intracellular part of the CD3ϵ chain, and phosphorylation of the immuno-receptor tyrosine-based activation motifs (ITAMs) in the intracellular domains of the CD3$\gamma\epsilon$, CD3$\delta\epsilon$, and CD3$\zeta\zeta$ dimers, which in turn lead to initiation of downstream signaling cascades and T cell activation.

It has long been recognized that the TCR–CD3 complex forms clusters upon ligand binding (31–36). More recently, it has been shown that in the absence of stimulating pMHC ligand, TCR–CD3 complexes are already expressed at the cell surface as a combination of monomeric and oligomeric TCR complexes or TCR nano-clusters (18). Electron microscopy (EM) analysis of immuno-gold-labeled human and murine T cells showed that these nano-clusters consist of up to 20 TCR–CD3 complexes. The exact stoichiometry of the nano-clusters has not been resolved yet.

The integrity of TCR nano-clusters depends on cholesterol present at the cell surface membrane (18). The formation of the clusters depends, at least, on the trans-membrane region of the CD3$\zeta\zeta$ homo-dimer (17), perhaps due to the capacity of $\zeta\zeta$ dimers to form dimers of dimers (37). Other possible mechanisms of cluster formation rely on the capacity of the extracellular domain of TCRα to dimerize (38).

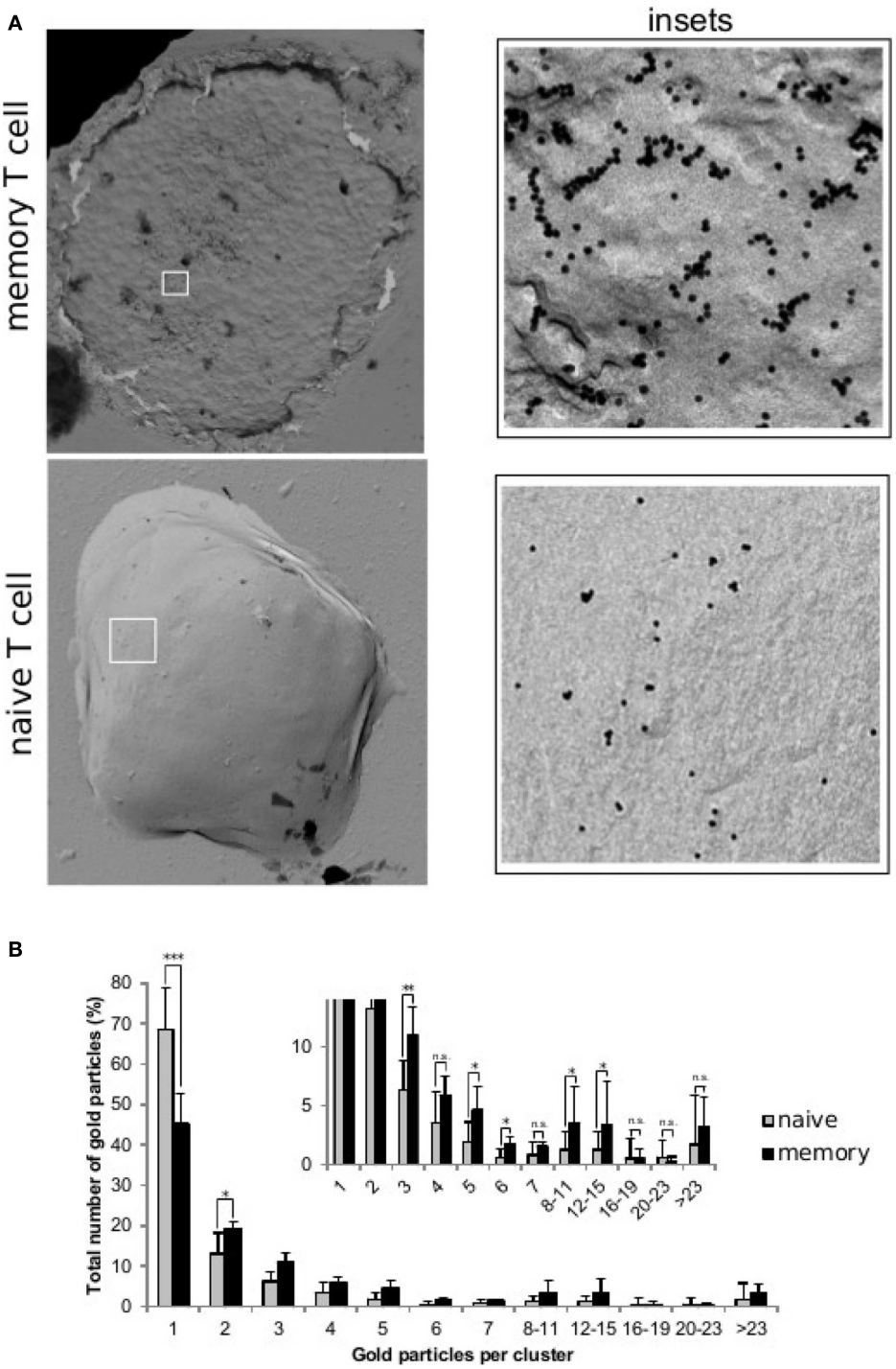


FIGURE 1 | Distribution of TCRs at the surface of naive and memory T cells. Resting naive and memory CD8$^+$ OT-1 T cells were labeled with the CD3ϵ-specific mAb 2C11 and 10 nm gold-conjugated protein-A. Cell surface replicas of the labeled T cells were analyzed by transmission electron microscopy and the number and size of the observed gold clusters were recorded. **(A)** TEM image of surface replicas of a memory and a naive OT-1 T cell. The insets to the right show an enlargement of the boxed areas. **(B)** Quantification (mean ± SD) of gold particles in clusters of the indicated sizes for resting naive T cells (gray bars, 7 cells, 9190 particles) and memory T cells (black bars, 5 cells, 3001 particles). The inset shows a detailed view of the distribution of clusters of three or more gold particles and statistical analysis (2-tailed Student's *t*-test: $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$). All naive and memory T cells had clusters with gold. However, whereas in naive T cells the maximum gold cluster size shared by all cells was four, this was eight for memory T cells. Also clusters bigger than twenty three particles were present in four out of five memory T cells, and only two out of seven naive T cells.

This body of experimental evidence allows us to conclude that multimeric TCR–CD3 complexes are co-expressed with TCR monomers on the surface of resting T cells.

A simple model of aggregation of TCR$\alpha\beta$ units is depicted in the left panel of **Figure 2**. Given a chain of length n (with n heterodimers linked), in a small time interval Δt, with probability $q_+ \Delta t$, the chain increases to length $n+1$, and with probability $q_- \Delta t$, the chain decreases to length $n-1$. Thus, by probability conservation, the probability to remain the same length n is $1-(q_+ + q_-)\Delta t$.

Mathematically, the dynamics of the process can be described by a continuous time Markov chain (39) (or birth and death process, as we assume that polymerization takes place in unit steps). The state space is $\{1, 2, 3, \ldots, n-1, n, n+1, \ldots\}$, where the number denotes the number of TCRs in a cluster:

$$1 \underset{q_+}{\overset{q_-}{\rightleftharpoons}} 2 \underset{q_+}{\overset{q_-}{\rightleftharpoons}} 3 \underset{q_+}{\overset{q_-}{\rightleftharpoons}} \cdots \underset{q_+}{\overset{q_-}{\rightleftharpoons}} n-1 \underset{q_+}{\overset{q_-}{\rightleftharpoons}} n \underset{q_+}{\overset{q_-}{\rightleftharpoons}} n+1 \cdots .$$

The forward Kolmogorov equations for the probability of having a cluster of size n are given by (40)

$$\frac{dp_n(t)}{dt} = q_+ p_{n-1}(t) + q_- p_{n+1}(t) - (q_+ + q_-)p_n(t)\,, \ \forall n \geq 2,$$
$$\frac{dp_1(t)}{dt} = q_- p_{2(t)} - q_+ p_1(t).$$

The stationary probability distribution is then given by

$$\lim_{t\to+\infty} p_n(t) \equiv \pi_n = \frac{b^{n-1}(1-b)}{(1-b^{N_{\max}})}\,,$$
$$b < 1\,, \quad n \in \{1, 2, 3, \ldots, N_{\max}\}\,, \qquad (1)$$

with $b = \frac{q_+}{q_-}$, and π_n the probability (in thermodynamic equilibrium) to have a cluster of size n. When $b<1$ (the number of clusters with a given size, n, decreases as n increases), and taking into account that peripheral T cells have around $N_{\max} \simeq 3 \times 10^4$ receptors, the latter expression can be further simplified to

$$\pi_n = b^{n-1}(1-b)\,, \quad b < 1\,, \quad n \in \{1, 2, 3, \ldots\}\,. \qquad (2)$$

2.2. MODEL 2: A BIVALENT MODEL FOR T CELL ACTIVATION

The TCR–pMHC binding model introduced in Ref. (26) considered monovalent pMHC ligands binding to TCR monomers on the surface of a T cell. Monovalent ligands have been reported to elicit a T cell response (41–43), but only when they are immobilized on a surface (which makes it difficult to assess whether they are truly monovalent or not). Yet, multivalent receptor–ligand interactions are required to elicit T cell responses in both $CD4^+$ and $CD8^+$ T cells. In what follows, and supported by a body of experimental work (3, 24, 44), we adopt the hypothesis that the minimum activating unit is a TCR–pMHC cross-linked dimeric complex (31, 45–47). We make use of the binding model (Model 2) with pMHC dimers (ligands) and dimeric TCRs (receptors), described in the right panel of **Figure 2**.

Gold-labeling experiments support the existence of nano-clusters with more than two TCRs, yet it can be shown (see Section 5.2) that the key parameter of the mathematical model is the fraction of monomeric to multimeric TCR clusters. Thus, without loss of generality, we will assume that all TCR clusters are dimeric.

The biochemical reactions encoded by the right panel of **Figure 2** are as follows:

- A (bivalent) ligand can bind a free receptor with monomeric binding reaction rates (k_{on} and k_{off}). Although not shown in the figure, we allow for a second ligand to bind the free receptor of the cluster. However, at low concentrations of ligands, this reaction can be safely neglected.
- Cross-linking of a singly bound ligand follows with rates k_2 (forward reaction) and k_{-2} (backward reaction).
- If the complex formed by the ligand cross-linked to the dimeric TCR cluster lasts at least a time τ, *dwell time*, we count that event. When we reach N such events, we will assume that a T cell response is initiated. The rationale behind this T cell response criterion follows the work of Palmer et al. (48), where the concepts of minimum dwell time and productive binding were introduced. This model combines aspects of the kinetic proof-reading (8) and the serial triggering models (7, 11). The

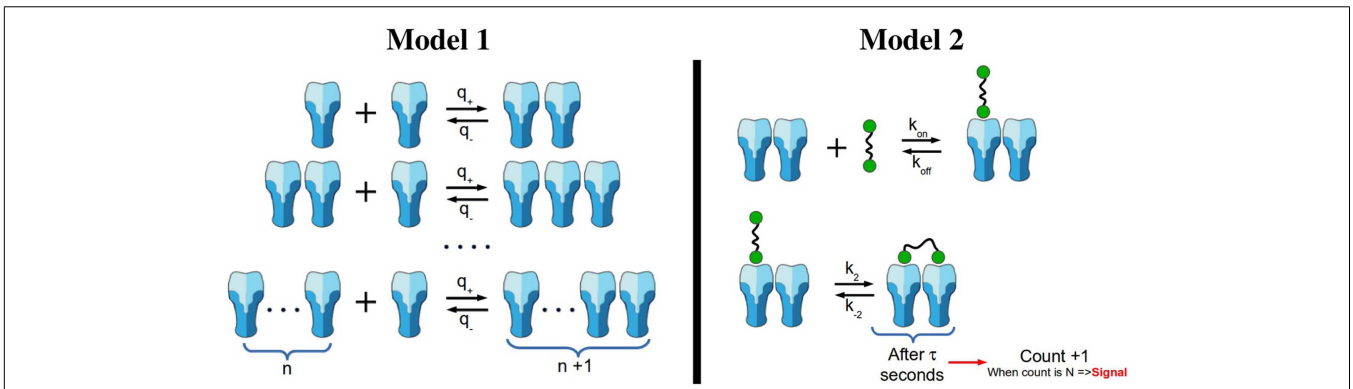


FIGURE 2 | Oligomerization and signaling models. Left panel: oligomerization model not mediated by ligand (Model 1). We assume that receptors are able to diffuse and aggregate to an existing cluster. However, we exclude the possibility of clusters with size larger than one to diffuse. Clusters grow one monomeric unit at a time. Right panel: reactions included in the stochastic activation model (Model 2). Ligands in solution are able to attach monovalently to any receptor in a cluster (top reaction). In addition, ligand-induced TCR cross-linking can occur once a ligand is bound to a TCR in a given nano-cluster (bottom reaction). Following Ref. (26), once the bivalently bound ligand has been attached for a time τ, we count that state as a signaling unit. After N of these units have been generated, the cell becomes activated.

minimum dwell time for a TCR–pMHC complex is the time the complex must remain bound in order to reach a level of ITAM phosphorylation, which will allow TCR triggering. Any binding, which persists for longer than the minimum dwell time is classified as a productive binding [see Refs. (48) and (26) for further details].

- From an immunological perspective, the relevant parameter is the *mean time to signal initiation*, or MTSI (26). Namely, the MTSI is the average time needed for a T cell response according to the criterion that at least N TCR dimers should be bivalently engaged to a bivalent ligand (pMHC) for at least a time τ.

Here we assume that N is around 10–100. That is, 10–100 TCRs are required for signaling and $N_B = b \times N_R$ is of the order of 10^4, with N_B the total number of clusters on the T cell surface. This means, under the assumptions of Model 2, that at most, there can be $N = 100$ internalization events, as this is the number of triggered TCRs. Thus, in this approximation, the loss of TCR due to internalization after triggering can be safely neglected. Nevertheless, internalization is an important step in early signaling, and a proper mechanistic model to justify the value of τ will require internalization to be considered. This analysis is out of the scope of this article.

We implement these reactions as a Markov process, and solve them numerically using the standard Gillespie algorithm (49), and with the parameters summarized in **Table 1**. We have made use of three different ligands: 4A, 4P, and 4N, which were also used in Ref. (26). For these ligands, that bind the same TCR with different affinities, a simple estimation of the number of cross-linking events required to elicit a T cell response is summarized in **Table 2**.

There is some evidence that, under physiological conditions, the chance of two specific peptides being presented by two MHC molecules in sufficient proximity and long enough to act as a dimer is very small (46). This will make ligand-induced TCR cross-linking a rare event. However, some recent experimental work on the distribution of cognate pMHC molecules on the surface of APCs shows that both for MHC class I (virus infection models), and for MHC class II (antigen uptake via the endocytic route) clusters of cognate pMHC can be detected (19–21).

We also note that ligand concentration is not the only factor that depends on physiological conditions. According to the DRiP hypothesis (28, 29), rapid viral antigen presentation is possible because antigenic peptides originate from defective ribosomal products that have short half-lives. Although this phenomenon affects the time between viral challenge and antigen presentation, we assume it is independent of the subsequent signaling dynamics of T cell activation.

Table 1 | Summary of the parameters used in the stochastic simulations.

Parameter	Value	Comment
N_A	6.023×10^{23}	Avogadro's number
N_R	30,000	Average number of TCRs per T cell (34)
V	50 μl	Volume of the experiment
N_C	10^5 cells	Number of T cells in the experiment
V_C	V/N_C	Average extracellular volume per cell
k_{-2}	k_{off}	Cross-linking off rate
k_2	$k_{off}(k_d/k_d^{\text{dimer}})$	Cross-linking rate[a]
N	10	Minimum number of bound dimer-bivalent clusters to elicit a T cell response
τ	1–4 s	Dwell time

For typical values of the dissociation rate, k_d, we find that k_2 is about 10–50 times k_{off}. We have assumed $k_{-2} = k_{off}$ following Ref. (44). When not explicitly shown, we have used the same parameters as in Ref. (26).

[a]The cross-linking rate k_2 is adapted from Ref. (44) for bivalent receptors.

Table 2 | Estimated mean number of cross-linking events, $N' \simeq N e^{2k_{-2}\tau}$, required to elicit a T cell response (SP thymocytes).

Ligand		N'	
	τ (s)	$N = 10$	$N = 100$
4P	1	3	12
($k_{\text{on}} = 153{,}691\ \text{M}^{-1}\ \text{s}^{-1}$)	4	3	13
($k_{\text{off}} = 0.0169\ \text{s}^{-1}$)	8	3	14
4A	1	7	58
($k_{\text{on}} = 157{,}533\ \text{M}^{-1}\ \text{s}^{-1}$)	4	$\sim 10^3$	$\sim 10^4$
($k_{\text{off}} = 0.8664\ \text{s}^{-1}$)	8	$\sim 10^6$	$\sim 10^7$
4N	1	$\sim 10^7$	$\sim 10^8$
($k_{\text{on}} = 149{,}385\ \text{M}^{-1}\ \text{s}^{-1}$)	4	$\sim 10^{30}$	$\sim 10^{31}$
($k_{\text{off}=8.6643}\ \text{s}^{-1}$)	8	$\sim 10^{60}$	$\sim 10^{61}$

3. RESULTS

3.1. DISTRIBUTION OF TCR CLUSTERS

The mathematical model described in Section 2.1, or Model 1, allows us to obtain the value of b that best fits the experimental data. We have used a weighted (by the variance) minimum-square regression to fit the experimental distributions to equation (2). This kind of fit minimizes the value of χ^2. Thus, in **Figure 3**, we show the agreement between theory and experiment, with values: $b_{\text{naive}} = 0.32$ and $b_{\text{memory}} = 0.55$. The difference between b_{naive} and b_{memory} can be explained by the existence of larger (or at least more localized) lipid rafts on the membrane of memory T cells (50, 51). Thus, the rates $q_{\pm}$ could be the effective combination of two mechanisms: one related to the diffusion of receptors on the membrane, and the other related to the aggregation of the receptors at the molecular level. The presence of cholesterol on the membrane changes the diffusion coefficient of the TCR receptors, as receptor diffusion within the raft is inhibited due to protein anchorage (52) and, thus, stabilizes the formation of clusters (a larger value of b means that, once two receptors are embedded in the same lipid raft, it is more difficult for them to become separated from each other).

A consequence of Model 1 is that, as the stationary probabilities need to sum up to one, the fraction of clusters of size larger than one is, precisely, b. This fraction is 72% higher for memory T cells than for naive T cells: $b_{\text{memory}}/b_{\text{naive}} = 1.72$.

3.2. MEAN TIME TO SIGNAL INITIATION

In **Figures 4A–D**, we show how the stochastic criterion is able to provide a ligand hierarchy according to their *potency*. Namely,

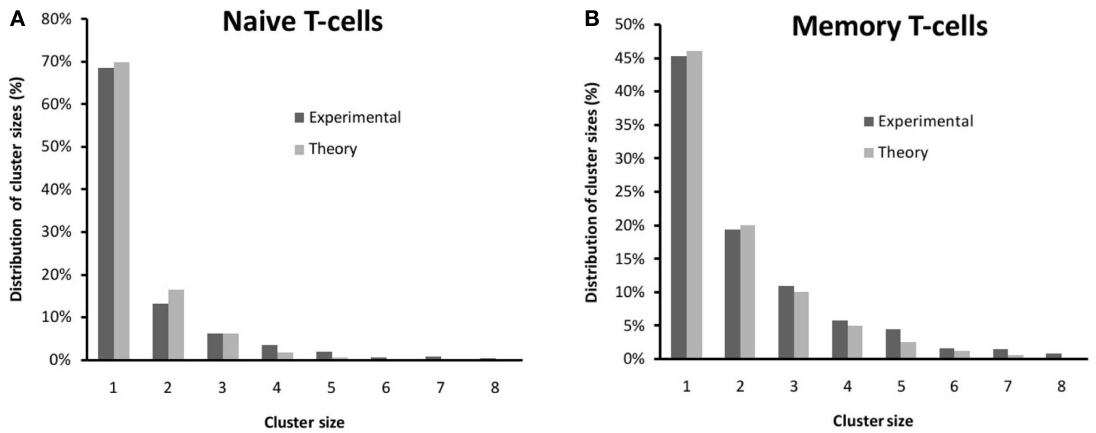


FIGURE 3 | Comparison between the experimental distribution of clusters (see also Figure 1) and those from Model 1 for (A) naive T cells and (B) memory T cells. The theoretical distribution has been fitted to equation (2) using a weighted (by the variance) minimum-square regression. The fitted values are $b_{naive} = 0.32$ and $b_{memory} = 0.55$.

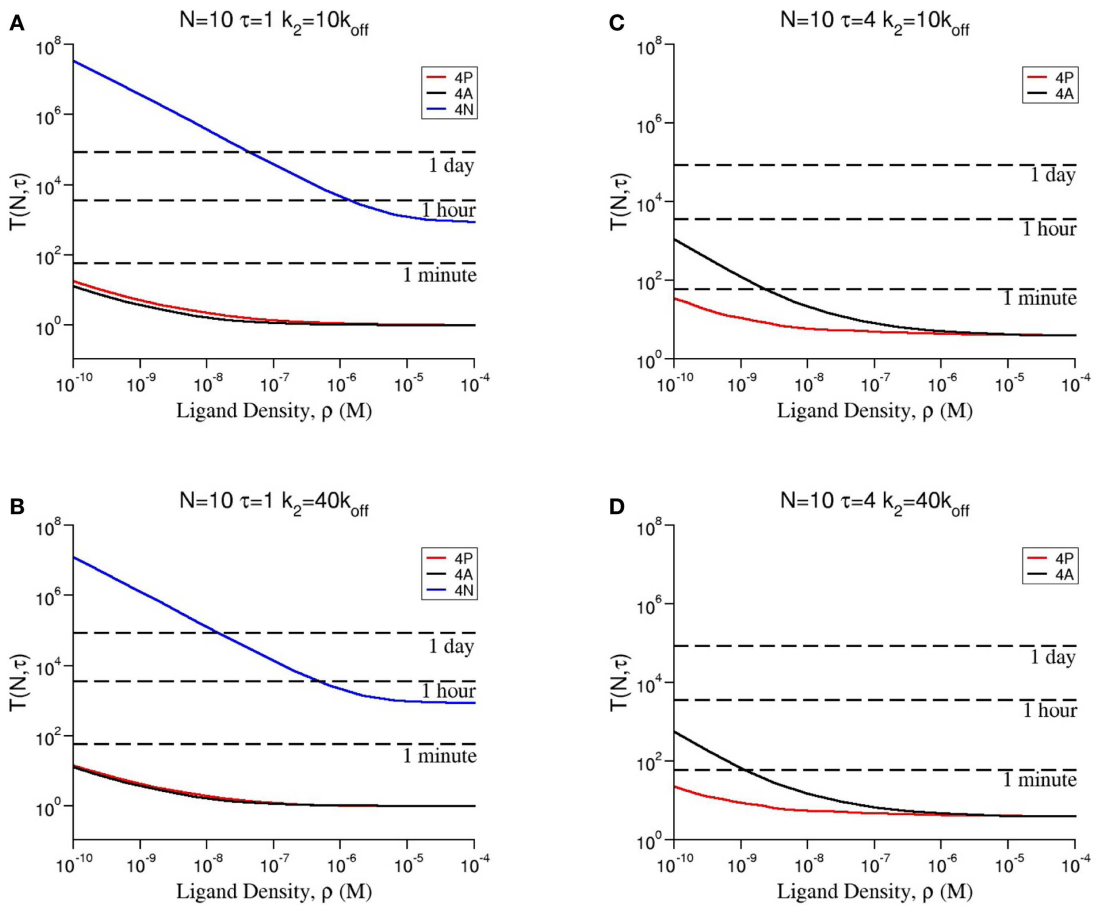


FIGURE 4 | Dependence of the mean time to signal initiation (MTSI), $T(N, \tau)$ to have N cross-linked ligand–receptor complexes bound for at least a dwell time τ for different model parameters as shown in every panel. The results have been obtained by making use of a Gillespie algorithm, after averaging over 100 realizations for each set of the parameters, summarized in **Table 2** (a python code for the stochastic integration is available upon request). Units of time are seconds. All parameters are taken from **Tables 1** and **2** except **(A)** $N = 10$, $\tau = 1$ and $k_2 = 10 \times k_{off}$; **(B)** $N = 10$, $\tau = 1$ and $k_2 = 40 \times k_{off}$; **(C)** $N = 10$, $\tau = 4$ and $k_2 = 10 \times k_{off}$; **(D)** $N = 10$, $\tau = 4$ and $k_2 = 40 \times k_{off}$.

the most agonistic ligand, 4P, elicits a T cell response in times of the order of a few seconds in all cases. On the contrary, the most antagonistic ligand, 4N, takes extremely large times to do so (in practical terms, this means it does not elicit a T cell response). Thus, TCR clustering can enhance the *potency* of ligands, when compared to the monomeric case (26), as experimentally observed and theoretically shown.

Following a similar approach to that of Ref. (26), we can derive an approximate formula for the mean time to signal initiation (MTSI), $T(N,\tau)$, for different ranges of ligand concentration, ρ. We write (see **Figure 5A** and Section 5.3 for further details):

$$T(N,\tau) \simeq \begin{cases} \tau & \text{at high concentration} \\ \tau + \left[\frac{N\exp(2k_{-2}\tau)}{2\rho N_B k_{\text{on}} k_2}\right]^{1/2} & \text{at intermediate concentration} \\ \tau + \frac{N\exp(2k_{-2}\tau)}{4\rho N_B k_{\text{on}}(k_2/k_{-2})} & \text{at low concentration} \end{cases} \quad (3)$$

These three regimes correspond to different immunological scenarios. In the case of high concentration of ligand, ligand is in great excess, so that the required number of *signaling units* is reached, almost as soon as the first signaling unit is formed (time of order τ). At low ligand concentration, the dynamics is limited by the first binding event, as cross-linking occurs in a slower timescale. So, the MTSI has the same functional form as that for the monomeric case (26). Finally, for intermediate ligand concentration, the competition between binding and cross-linking implies a more complicated mathematical relationship. Of greater relevance to the discussion is the nature of the ligand (with different k_{on} and k_{off} rates), and the number of TCR clusters on the membrane of the T cell (encoded in the parameter $N_B = b \times N_R$, with N_R, the average number of TCRs per T cell, see **Table 1**).

An expression for the variance of the time to signal initiation (TSI) cannot be provided in a closed form [as done in Ref. (26)]. However, the fact that the variance decreases as the ligand concentration increases, suggests that the mathematical formula for the variance in the monovalent case can provide an upper bound to the present (dimeric) case.

Using equation (3), we also can deduce the role of pre-clustering in the signaling time, or MTSI. As the number of bivalent clusters is $b \times N_R$, the larger b is, the shorter the response time becomes. The model predicts that, for physiological conditions (not too high ligand concentrations), the ratio of the MTSI for naive and memory T cells is inversely proportional to the ratio of their corresponding values of b. Namely, memory cells would respond up to 72% faster than naive ones (**Figure 5B**).

4. DISCUSSION

TCR triggering mechanisms are currently under debate [see, for example, Ref. (53) and (7) for recent reviews]. TCR clustering may be invoked as a description of the experimental results (27). The requirement for multivalent engagement of TCRs by pMHC ligands in $CD4^+$ T cells has been widely shown (45, 47, 54, 55). The same requirement was shown in $CD8^+$ T cells by Stone and Stern (56).

In this paper, we have made use of the concept of *mean time to signal initiation* (MTSI or stochastic criterion) as a method to quantify the effect of TCR clustering on the timescales of T cell responses and, thus, to compare the behavior of naive and memory T cells. This criterion has also allowed us to compare the results in Section 3 for dimeric binding with those of Ref. (26) for monomeric binding. The introduction of the cross-linked ligand–receptor complex as the minimum *signaling* unit gives the response greater sensitivity to small differences in ligand affinity.

A recent and novel feature of TCR immunology is the existence of TCR nano-clusters that are pre-formed, independently of ligand (3). This suggests that a simple stoichiometric clustering model (oligomerization of free TCRs diffusing on the T cell membrane) is enough to account for the distribution of TCR nano-clusters. In the case of naive T cells, Model 1 predicts an effective non-dimensional parameter, $b = q_+/q_-$, that allows us to explain the experimentally observed TCR cluster distributions. The presence of larger lipid rafts on the membrane of memory T cells might provide support for the different values of b for naive and memory cells, b_{naive} and b_{memory}, respectively. It has recently been shown that receptor diffusion within the raft is inhibited due to protein anchorage (52). This reduction in the TCR diffusion coefficient

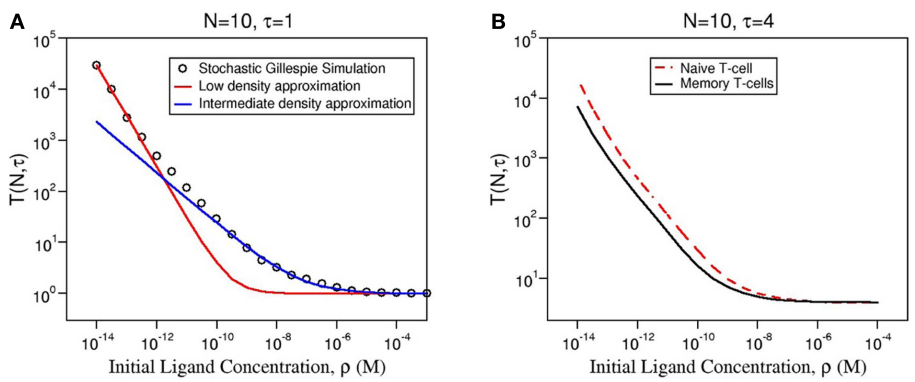


FIGURE 5 | (A) Comparison of the numerical solution of Model 2 (Gillespie algorithm with the parameters summarized in **Table 2**) and the approximate solution [equation (3)] for ligand 4P and the same parameters as in **Figure 4**. **(B)** Comparison of the mean MTSI for naive (red dashed line) and memory (solid black line) T cells.

would increase the time required for the receptor to escape from the raft [in a similar fashion as other escape problems (57)]. This escape time is inversely proportional to the diffusion coefficient itself. A smaller TCR diffusivity, as would be the case for memory T cells, will imply a larger residence time in the raft, which in turn will increase the probability of receptor aggregation in a given TCR cluster. A more detailed model of TCR diffusion and aggregation on the T cell membrane will be the subject of future work.

Equation (3) shows the explicit dependence of the MTSI, $T(N, \tau)$, on the parameter N_B, for given values of N and τ. N_B is the average number of dimeric receptor clusters per T cell, so that $N_B = b \times N_R$, with N_R, the average number of TCRs per T cell (see **Table 1**). For large ligand concentration, the predicted T cell response time for memory and naive T cells is the same, and is equal to τ. In the case of intermediate concentrations, the MTSI is proportional to $\frac{1}{\sqrt{b}}$. Finally, for low ligand concentration the MTSI is proportional to $\frac{1}{b}$. This implies that, at low ligand concentration, TCR pre-clustering alone can only account for at most 72% of the reduction in the response time between memory and naive T cells. This behavior is illustrated in **Figure 5A**. This difference is not so large as to be able to account for the observed higher responsiveness of memory T cells. Our results, thus, point to the need for additional mechanisms beyond TCR pre-clustering.

A potential candidate to explain the large differences between memory and naive T cell responses is the conformational change of the CD3 complex (58). This conformational change is essential to enable ITAM phosphorylation and, thus, the transfer of the TCR signal from the ecto-domain to the cytoplasmic tail of the TCR (58). Conformational changes in the CD3 complex occur as a result of the $\alpha\beta$ hetero-dimer binding to pMHC. These conformational changes allow the subunits of the CD3 complex (the $\gamma\epsilon$ and $\delta\epsilon$ hetero-dimers and the $\zeta\zeta$ homo-dimer) to become accessible to Lck, which can then phosphorylate their cytoplasmic domains at the ITAMs, leading to T cell signaling (59). In this way, the ligand-induced conformational change of the receptors can be propagated to all the receptors in the same cluster, so that larger clusters would benefit from this conformational change as a *cascade* [see, for example, Ref. (60) and references therein]. Thus, differences in the distribution of cluster sizes could, indeed, explain the immunological differences between memory and naive cells.

Other membrane receptors also exhibit pre-clustering and ligand-induced receptor cross-linking. For instance, in the case of the vascular endothelial growth factor receptor (VEGFR), it has been shown (61) that there are two distinct pathways to receptor dimerization: (i) dynamic pre-dimerization (as the one described in Model 1), and (ii) ligand-induced receptor dimerization. The main conclusion in Ref. (61) is that both mechanisms are almost indistinguishable at low ligand concentration. However, the first mechanism is more sensitive to changes in the binding affinity at large ligand concentration. Although the biological system studied in Ref. (61) is different from the T cell receptor considered here, their conclusions might be generalized as both receptors are tyrosine kinases.

Bachmann et al. (62, 63) considered a model of diffusion and ligand-induced TCR clustering. Their model suggests that the existence of large enough clusters greatly inhibits subsequent multimer diffusion, thus, reducing the relevance that this mechanism might have. This inhibition might be experimentally tested by exploiting the differences between naive (small and few clusters) and memory (large and many clusters). It will be interesting to make use of the models introduced in this paper to investigate the different roles of ligand binding and cellular activation (62), and TCR turnover (64).

Finally, the existence of TCR pre-clusters [and the knowledge of their membrane distribution given by π_n, equation (2)] can be considered in the kinetic-segregation model (65). In this model, diffusion out of close-contact zones would be inhibited by the existence of nano-clusters, thus, enhancing the number of triggered receptors. In a similar way, consecutive receptor phosphorylation events (66) in TCR nano-clusters would also amplify receptor signaling.

5. MATERIALS AND METHODS

5.1. EXPERIMENTS

Naive $CD8^+$ OT-1 T cells, which recognize an ovalbumin-derived peptide presented by the MHC class I molecule H-$2K^b$, were isolated from superficial and mesenteric lymph nodes of OT-1 TCR transgenic mice (67), via depletion of $CD19^+$ B cells, $CD4^+$ helper T cells and $CD11b^+$ macrophages, using antibodies and Dynal magnetic beads (Invitrogen). Memory OT-1 T cells were generated by adoptively transferring 10^6 naive OT-1 T cells into congenic C57BL/6 Ly5.1 Pep3b mice, which were simultaneously immunized with 10^7 PFU MVA-OVA (68). After 6 months, resting memory OT-1 T cells were isolated from the spleen and lymph nodes of these mice by antibody-mediated depletion of macrophages, B cells, and $CD4^+$ T cells, followed by separation of the OT-1 memory T cells from host-derived $Ly5.1^+$ $CD8^+$ T cells via fluorescence-activated cell sorting, using a Ly5.1-specific antibody. Labeling of cells with the CD3ϵ-specific antibody 2C11 and 10 nm gold-conjugated protein-A, replica generation and analysis were performed as previously described (17).

5.2. MODELS OF SIGNALING WITH DIMERIC AND TRIMERIC RECEPTOR CLUSTERS

In Section 2, we introduced a model in which ligands are bivalent and receptor clusters are dimeric (that is, composed of two monomeric TCRs). This is, of course, a first approximation that neglects the distribution of cluster sizes experimentally observed. Yet, the results of our mathematical study only change in a quantitative way, but not qualitatively, when we include TCR clusters of larger sizes. In this Section, we illustrate this by considering a system in which clusters of size 1, 2, and 3 coexist and the ligands are bivalent. **Table 3** provides the notation introduced to describe the molecular species considered in the model, as well as a graphical representation.

At large initial ligand concentration, under the stochastic criterion, the MTSI tends to τ. On the other hand, at low initial ligand concentration, the number of receptors, compared to the number of ligands, is so large that we can neglect molecular species x_4, y_4, y_6, and y_7, which involve more than one bivalent ligand. This has also been confirmed experimentally. Given our stochastic T cell response criterion, in this case, the signaling units correspond to molecular species x_5, y_5, and y_7. Molecular species z_1 and z_3 do

Table 3 | Summary of variables for a model in which clusters of size 1, 2, and 3 coexist.

Variable	Description	Molecular representation
z_1	Free monomeric receptor	
z_2	Free ligand (dimer)	
z_3	Ligand-bound to a monomeric receptor	
x_1	Free dimeric cluster	
x_2	Same as z_2 (defined for convenience of notation)	
x_3	Ligand singly bound to a dimeric cluster	
x_4	Two ligands bound to a dimeric cluster	
x_5	Cross-linked ligand in a dimeric cluster	
y_1	Free trimeric receptor	
y_2	Same as z_2 (defined for convenience of notation)	
y_3	Ligand singly bound to a trimeric cluster	
y_4	Two ligands bound to a trimeric cluster	
y_5	Cross-linked ligand in a trimeric cluster	
y_6	Three ligands bound to a trimeric cluster	
y_7	One ligand singly bound to a trimeric cluster and another cross-linked	

All the variables correspond to the total number of molecular species (not concentrations). Hence, all the rates in the mathematical model have units of s^{-1}.

not contribute to the T cell response and will be neglected in what follows. Thus, we only need to consider the dynamics of dimeric and trimeric T cell receptor clusters.

We introduce the total number of *signaling units*, $S_5(t) \equiv x_5(t) + y_5(t)$. The set of ordinary differential equations for the model is given by:

$$\begin{aligned}
\dot{x}_1 &= -4k_+x_1x_2 + k_{\text{off}}x_3,\\
\dot{x}_2 &= -4k_+x_1x_2 + k_{\text{off}}x_3,\\
\dot{x}_3 &= 4k_+x_1x_2 - k_{\text{off}}x_3 - k_2x_3 + 2k_{-2}x_5,\\
\dot{x}_5 &= k_2x_3 - 2k_{-2}x_5,\\
\dot{y}_1 &= -6k_+y_1y_2 + k_{\text{off}}y_3,\\
\dot{y}_2 &= -6k_+y_1y_2 + k_{\text{off}}y_3,\\
\dot{y}_3 &= 6k_+y_1y_2 - k_{\text{off}}y_3 - 2k_2y_3 + 2k_{-2}y_5,\\
\dot{y}_5 &= 2k_2y_3 - 2k_{-2}y_5,
\end{aligned}$$

where $k_+ = k_{\text{on}}/(VN_A)$, V is the volume of the experiment and N_A is Avogadro's number.

Given the symmetry of the problem, and in the limit of low initial ligand concentration, we will assume that the ratio of x_3 to y_3 is that of the initial ratio of free TCR dimers to free TCR trimers, namely,

$$\frac{y_3}{x_3} \simeq \frac{\pi_3}{\pi_2} = b \quad \Rightarrow \quad y_3 \simeq b\, x_3\,, \tag{4}$$

where we have made use of equation (2) to conclude $\frac{\pi_3}{\pi_2} = b$. Thus, the total number of signaling units, $S_5(t)$, obeys the following differential equation

$$\dot{S}_5 = \dot{x}_5 + \dot{y}_5 = k_2\,(1+2b)\,x_3 - 2\,k_{-2}\,S_5\,. \tag{5}$$

Finally, in the low ligand concentration limit as above, let us introduce $S_3 \equiv x_3 + y_3$. It is easy to show that equation (5) reduces to

$$\dot{S}_5 = k_2\,\frac{1+2b}{1+b}\,S_3 - 2\,k_{-2}\,S_5\,, \tag{6}$$

which is identical to the differential equation for x_5 above, but with $S_{5,3}$ replaced by $x_{5,3}$, respectively. This means that, except for a prefactor $\frac{1+2b}{1+b}$ [which, for $b \in (0, 1)$, is between 1 and 3/2], the study of dimeric and trimeric clusters is reduced to the dimeric case.

5.3. A SIMPLE FORMULA FOR THE MTSI

The basic idea behind the stochastic criterion is to count the cumulative number of events that may contribute to signaling (26). Here, we calculate the mean number of cross-linking events up to time t, $C(t)$, as the integral,

$$C(t) = k_2 \int_0^t x_3(s)\, ds\,. \tag{7}$$

It is possible to obtain an expression for $x_3(t)$ with the approximation that the product $x_1(t)x_2(t)$ is constant, so that the pair of equations for $x_3(t)$ and $x_5(t)$ can be solved exactly. This yields (69):

$$\begin{aligned}
C(t) = k_2 \Bigg[&\frac{c_1}{\lambda_1}(\lambda_1 + 2k_{-2})(e^{\lambda_1 t} - 1)\\
&+ \frac{c_2}{\lambda_2}(\lambda_2 + 2k_{-2})(e^{\lambda_2 t} - 1) + a_1 t \Bigg]\,,
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
c_1 &= \frac{-4\lambda_2 k_{\text{on}}\rho N_B}{(\lambda_2 - \lambda_1)(4k_{\text{on}}\rho k_2 + 2k_{\text{off}}k_{-2} + 8k_{\text{on}}\rho k_{-2})}\,,\\
c_2 &= \frac{4\lambda_1 k_{\text{on}}\rho N_B}{(\lambda_2 - \lambda_1)(4k_{\text{on}}\rho k_2 + 2k_{\text{off}}k_{-2} + 8k_{\text{on}}\rho k_{-2})}\,,\\
\lambda_{1,2} &= \frac{1}{2}\,(-4k_{\text{on}}\rho - k_{\text{off}} - k_2 - 2k_{-2}\\
&\quad \pm \left[(4k_{\text{on}}\rho + k_{\text{off}} - k_2 - 2k_{-2})^2 + 4k_{\text{off}}k_2\right]^{1/2}\Big)\,,\\
a_1 &= \frac{8k_{-2}k_{\text{on}}\rho N_B}{4k_{\text{on}}\rho k_2 + 2k_{\text{off}}k_{-2} + 8k_{\text{on}}\rho k_{-2}}\,,\\
a_2 &= \frac{4k_2 k_{\text{on}}\rho N_B}{4k_{\text{on}}\rho k_2 + 2k_{\text{off}}k_{-2} + 8k_{\text{on}}\rho k_{-2}}\,,
\end{aligned}$$

and N_B is the number of dimeric receptors. The MTSI is then given by the solution of the equation $C(T(N, \tau)-\tau) = N \exp(2k_{-2}\tau)$.

The expressions in equation (3) are obtained from equation (8) in the appropriate regimes. At low ligand concentration, $C(t)$ is simply proportional to time: $C(t) \simeq k_2 a_1 t$, so that $C(T-\tau) = k_2 a_1 (T-\tau) = N \exp(2k_{-2}\tau)$. When $\lambda_{1,2}\tau \ll 1$, on the other hand, the first non-zero term in a Taylor expansion of $C(t)$ in time is quadratic: $C(t) \propto t^2$. This provides the exponent 1/2 in the second line of equation (3).

ACKNOWLEDGMENTS

We thank Ed Palmer, Wolfgang Schamel, and Thomas Höfer for helpful discussions. We also thank the Max Planck Institute for the Physics of Complex Systems (Dresden) and the International Centre for Mathematical Sciences (Edinburgh), where part of this work was discussed and presented, for their hospitality. This work has been partially supported through Grants No. FIS2009-12964-C05-03 (Mario Castro, Grant Lythe, Carmen Molina-París), BFU2009-08009 from the Ministerio de Ciencia e Innovación (Hisse M. van Santen), FP7 PIRSES-GA-2008-230665 and PIRSES-GA-2012-317893 (Mario Castro, Grant Lythe, and Carmen Molina-París), BBSRC BB/F003811/1 (Grant Lythe and Carmen Molina-París), and BBSRC BB/G023395/1 (Carmen Molina-París).

REFERENCES

1. Huppa JB, Davis MM. T-cell-antigen recognition and the immunological synapse. *Nat Rev Immunol* (2003) **3**(12):973–83. doi:10.1038/nri1245
2. Batista FD, Dustin ML. Cell: cell interactions in the immune system. *Immunol Rev* (2013) **251**(1):7–12. doi:10.1111/imr.12025
3. Schamel WW, Alarcón B. Organization of the resting TCR in nanoscale oligomers. *Immunol Rev* (2013) **251**(1):13–20. doi:10.1111/imr.12019
4. Allard JF, Dushek O, Coombs D, van der Merwe PA. Mechanical modulation of receptor-ligand interactions at cell-cell interfaces. *Biophys J* (2012) **102**(6):1265–73. doi:10.1016/j.bpj.2012.02.006
5. Lillemeier BF, Mörtelmaier MA, Forstner MB, Huppa JB, Groves JT, Davis MM. TCR and Lat are expressed on separate protein islands on T cell membranes and concatenate during activation. *Nat Immunol* (2009) **11**(1):90–6. doi:10.1038/ni.1832
6. Edwards LJ, Zarnitsyna VI, Hood JD, Evavold BD, Zhu C. Insights into t cell recognition of antigen: significance of two-dimensional kinetic parameters. *Front Immunol* (2012) **3**:86. doi:10.3389/fimmu.2012.00086
7. Zarnitsyna V, Zhu C. T cell triggering: insights from 2D kinetics analysis of molecular interactions. *Phys Biol* (2012) **9**(4):045005. doi:10.1088/1478-3975/9/4/045005
8. McKeithan T. Kinetic proofreading in T-cell receptor signal transduction. *Proc Natl Acad Sci U S A* (1995) **92**(11):5042. doi:10.1073/pnas.92.11.5042
9. Kalergis AM, Boucheron N, Doucey M-A, Palmieri E, Goyarts EC, Vegh Z, et al. Efficient T cell activation requires an optimal dwell-time of interaction between the TCR and the pMHC complex. *Nat Immunol* (2001) **2**(3):229–34. doi:10.1038/85286
10. Matis LA, Glimcher LH, Paul WE, Schwartz RH. Magnitude of response of histocompatibility-restricted T-cell clones is a function of the product of the concentrations of antigen and IA molecules. *Proc Natl Acad Sci U S A* (1983) **80**(19):6019–23. doi:10.1073/pnas.80.19.6019
11. Valitutti S, Müller S, Cella M, Padovan E, Lanzavecchia A. Serial triggering of many T-cell receptors by a few peptide MHC complexes. *Nature* (1995) **375**(6527):148–51. doi:10.1038/375148a0
12. Friedl P, Gunzer M. Interaction of T cells with APCS: the serial encounter model. *Trends Immunol* (2001) **22**(4):187–91. doi:10.1016/S1471-4906(01)01869-5
13. Dushek O, Das R, Coombs D. A role for rebinding in rapid and reliable T cell responses to antigen. *PLoS Comput Biol* (2009) **5**(11):e1000578. doi:10.1371/journal.pcbi.1000578
14. Daniels MA, Teixeiro E, Gill J, Hausmann B, Roubaty D, Holmberg K, et al. Thymic selection threshold defined by compartmentalization of RAS/MAPK signalling. *Nature* (2006) **444**(7120):724–9. doi:10.1038/nature05269
15. Valitutti S, Coombs D, Dupré L. The space and time frames of T cell activation at the immunological synapse. *FEBS Lett* (2010) **584**(24):4851–7. doi:10.1016/j.febslet.2010.10.010
16. Smith-Garvin JE, Koretzky GA, Jordan MS. T cell activation. *Annu Rev Immunol* (2009) **27**:591. doi:10.1146/annurev.immunol.021908.132706
17. Kumar R, Ferez M, Swamy M, Arechaga I, Rejas MT, Valpuesta JM, et al. Increased sensitivity of antigen-experienced T cells through the enrichment of oligomeric T cell receptor complexes. *Immunity* (2011) **35**(3):375–87. doi:10.1016/j.immuni.2011.08.010
18. Schamel WW, Arechaga I, Risueño RM, van Santen HM, Cabezas P, Risco C, et al. Coexistence of multivalent and monovalent TCRs explains high sensitivity and wide range of response. *J Exp Med* (2005) **202**(4):493–503. doi:10.1084/jem.20042155
19. Lu X, Gibbs JS, Hickman HD, David A, Dolan BP, Jin Y, et al. Endogenous viral antigen processing generates peptide-specific MHC class I cell-surface clusters. *Proc Natl Acad Sci U S A* (2012) **109**(38):15407–12. doi:10.1073/pnas.1208696109
20. Ferez M, Castro M, Alarcon B, van Santen HM. Cognate peptide-MHC complexes are expressed as tightly apposed nanoclusters in virus-infected cells to allow tcr crosslinking. *J Immunol* (2014) **192**(1):52–8. doi:10.4049/jimmunol.1301224
21. Bosch B, Heipertz EL, Drake JR, Roche PA. Major histocompatibility complex (MHC) class II-peptide complexes arrive at the plasma membrane in cholesterol-rich microclusters. *J Biol Chem* (2013) **288**(19):13236–42. doi:10.1074/jbc.M112.442640
22. Goldstein B, Faeder JR, Hlavacek WS. Mathematical and computational models of immune-receptor signalling. *Nat Rev Immunol* (2004) **4**(6):445–56. doi:10.1038/nri1374
23. Alarcón B, Swamy M, van Santen HM, Schamel WW. T-cell antigen-receptor stoichiometry: pre-clustering for sensitivity. *EMBO Rep* (2006) **7**(5):490–5. doi:10.1038/sj.embor.7400682
24. Bray D, Levin MD, Morton-Firth CJ. Receptor clustering as a cellular mechanism to control sensitivity. *Nature* (1998) **393**(6680):85–8. doi:10.1038/30018
25. Slifka MK, Whitton JL. Functional avidity maturation of $CD8^+$ T cells without selection of higher affinity TCR. *Nat Immunol* (2001) **2**(8):711–7. doi:10.1038/90650
26. Currie J, Castro M, Lythe G, Palmer E, Molina-París C. A stochastic T cell response criterion. *J R Soc Interface* (2012) **9**(76):2856–70. doi:10.1098/rsif.2012.0205
27. van der Merwe P, Dushek O. Mechanisms for T cell receptor triggering. *Nat Rev Immunol* (2010) **11**(1):47–55. doi:10.1038/nri2887
28. Yewdell JW. Drips solidify: progress in understanding endogenous MHC class I antigen processing. *Trends Immunol* (2011) **32**(11):548–58. doi:10.1016/j.it.2011.08.001
29. Rock KL, Farfán-Arribas DJ, Colbert JD, Goldberg AL. Re-examining class-I presentation and the DRiP hypothesis. *Trends Immunol* (2014). doi:10.1016/j.it.2014.01.002
30. Antón LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J Leukoc Biol* (2014). doi:10.1189/jlb.1113599
31. Stone J, Cochran J, Stern L. T-cell activation by soluble MHC oligomers can be described by a two-parameter binding model. *Biophys J* (2001) **81**(5):2547–57. doi:10.1016/S0006-3495(01)75899-7
32. Stone J, Chervin A, Kranz D. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology* (2009) **126**(2):165–76. doi:10.1111/j.1365-2567.2008.03015.x
33. Coombs D, Dushek O, Merwe P. A review of mathematical models for T cell receptor triggering and antigen discrimination. In: Molina-París C, Lythe G, editors. *Mathematical Models and Immune Cell Biology*. New York: Springer (2011). p. 25–45.
34. Coombs D, Kalergis AM, Nathenson SG, Wofsy C, Goldstein B. Activated TCRs remain marked for internalization after dissociation from pMHC. *Nat Immunol* (2002) **3**(10):926–31. doi:10.1038/ni838
35. Choudhuri K, Dustin ML. Signaling microdomains in T cells. *FEBS Lett* (2010) **584**(24):4823–31. doi:10.1016/j.febslet.2010.10.015

36. Yokosuka T, Saito T. The immunological synapse, TCR microclusters, and T cell activation. In: Saito T, Batista FD, editors. *Immunological Synapse*. Berlin Heidelberg: Springer (2010). p. 81–107.
37. Torres J, Briggs JA, Arkin IT. Multiple site-specific infrared dichroism of CD3-ζ, a transmembrane helix bundle. *J Mol Biol* (2002) **316**(2):365–74. doi:10.1006/jmbi.2001.5267
38. Kuhns MS, Girvin AT, Klein LO, Chen R, Jensen KD, Newell EW, et al. Evidence for a functional sidedness to the $\alpha\beta$TCR. *Proc Natl Acad Sci U S A* (2010) **107**(11):5094–9. doi:10.1073/pnas.1000925107
39. Norris JR. *Markov Chains*. Cambridge University Press (1998).
40. Taylor H, Karlin S. *An Introduction to Stochastic Modeling*. San Diego: Academic Press (1998).
41. Janeway CA Jr. Ligands for the T-cell receptor: hard times for avidity models. *Immunol Today* (1995) **16**(5):223–5. doi:10.1016/0167-5699(95)80163-4
42. Ma Z, Sharp KA, Janmey PA, Finkel TH. Surface-anchored monomeric agonist pMHCs alone trigger TCR with high sensitivity. *PLoS Biol* (2008) **6**(2):e43. doi:10.1371/journal.pbio.0060043
43. Huang J, Brameshuber M, Zeng X, Xie J, Li QJ, Chien YH, et al. A single peptide-major histocompatibility complex ligand triggers digital cytokine secretion in CD4+ T cells. *Immunity* (2013) **39**:846–57. doi:10.1016/j.immuni.2013.08.036
44. Stone J, Artyomov M, Chervin A, Chakraborty A, Eisen H, Kranz D. Interaction of streptavidin-based peptide-MHC oligomers (tetramers) with cell-surface TCRs. *J Immunol* (2011) **187**(12):6281–90. doi:10.4049/jimmunol.1101734
45. Abastado J-P, Lone Y-C, Casrouge A, Boulot G, Kourilsky P. Dimerization of soluble major histocompatibility complex-peptide complexes is sufficient for activation of T cell hybridoma and induction of unresponsiveness. *J Exp Med* (1995) **182**(2):439–47. doi:10.1084/jem.182.2.439
46. Cochran JR, Cameron TO, Stern LJ. The relationship of MHC-peptide binding and T cell activation probed using chemically defined MHC class II oligomers. *Immunity* (2000) **12**(3):241–50. doi:10.1016/S1074-7613(00)80177-6
47. Boniface JJ, Rabinowitz JD, Wülfing C, Hampl J, Reich Z, Altman JD, et al. Initiation of signal transduction through the T cell receptor requires the multivalent engagement of peptide/MHC ligands. *Immunity* (1998) **9**(4):459–66. doi:10.1016/S1074-7613(00)80629-9
48. Palmer E, Naeher D. Affinity threshold for thymic selection through a T-cell receptor-co-receptor zipper. *Nat Rev Immunol* (2009) **9**(3):207–13. doi:10.1038/nri2469
49. Gillespie D. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* (1977) **81**:2340–61. doi:10.1021/j100540a008
50. Brumeanu T-D, Preda-Pais A, Stoica C, Bona C, Casares S. Differential partitioning and trafficking of GM gangliosides and cholesterol-rich lipid rafts in thymic and splenic CD4 T cells. *Mol Immunol* (2007) **44**(4):530–40. doi:10.1016/j.molimm.2006.02.008
51. Kersh EN, Kaech SM, Onami TM, Moran M, Wherry EJ, Miceli MC, et al. TCR signal transduction in antigen-specific memory CD8 T cells. *J Immunol* (2003) **170**(11):5455–63.
52. Lingwood D, Simons K. Lipid rafts as a membrane-organizing principle. *Science* (2010) **327**(5961):46–50. doi:10.1126/science.1174621
53. Robert P, Aleksic M, Dushek O, Cerundolo V, Bongrand P, van der Merwe P. Kinetics and mechanics of two-dimensional interactions between T cell receptors and different activating ligands. *Biophys J* (2012) **102**(2):248–57. doi:10.1016/j.bpj.2011.11.4018
54. Cochran JR, Stern LJ. A diverse set of oligomeric class II MHC-peptide complexes for probing T-cell receptor interactions. *Chem Biol* (2000) **7**(9):683–96. doi:10.1016/S1074-5521(00)00019-3
55. Dushek O, Goyette J, Merwe PA. Non-catalytic tyrosine-phosphorylated receptors. *Immunol Rev* (2012) **250**(1):258–76. doi:10.1111/imr.12008
56. Stone JD, Stern LJ. CD8 T cells, like CD4 T cells, are triggered by multivalent engagement of TCRs by MHC-peptide ligands but not by monovalent engagement. *J Immunol* (2006) **176**(3):1498–505.
57. Day M, Lythe G. Timescales of the adaptive immune response. In: Molina-París C, Lythe G, editors. *Mathematical Models and Immune Cell Biology*. New York: Springer (2011). p. 351–61.
58. Minguet S, Swamy M, Alarcón B, Luescher IF, Schamel WW. Full activation of the T cell receptor requires both clustering and conformational changes at CD3. *Immunity* (2007) **26**(1):43–54. doi:10.1016/j.immuni.2006.10.019
59. Marks F, Klingmüller U, Müller-Decker K. *Cellular Signal Processing: An Introduction to the Molecular Mechanisms of Signal Transduction*. New York, NY: Garland Science (2009).
60. Blanco R, Alarcón B. TCR nanoclusters as the framework for transmission of conformational changes and cooperativity. *Front Immunol* (2012) **3**:115. doi:10.3389/fimmu.2012.00115
61. Mac Gabhann F, Popel AS. Dimerization of VEGF receptors and implications for signal transduction: a computational study. *Biophys Chem* (2007) **128**(2–3):125–39. doi:10.1016/j.bpc.2007.03.010
62. Bachmann MF, Salzmann M, Oxenius A, Ohashi PS. Formation of TCR dimers/trimers as a crucial step for T cell activation. *Eur J Immunol* (1998) **28**(8):2571–9. doi:10.1002/(SICI)1521-4141(199808)28:08<2571::AID-IMMU2571>3.0.CO;2-T
63. Bachmann MF, Ohashi PS. The role of T-cell receptor dimerization in T-cell activation. *Immunol Today* (1999) **20**(12):568–76. doi:10.1016/S0167-5699(99)01543-1
64. Sousa J, Carneiro J. A mathematical analysis of tcr serial triggering and down-regulation. *Eur J Immunol* (2000) **30**:3219–27. doi:10.1002/1521-4141(200011)30:11<3219::AID-IMMU3219>3.0.CO;2-7
65. Davis SJ, van der Merwe PA. The kinetic-segregation model: TCR triggering and beyond. *Nat Immunol* (2006) **7**(8):803–9. doi:10.1038/ni1369
66. Mukhopadhyay H, Cordoba S-P, Maini PK, van der Merwe PA, Dushek O. Systems model of T cell receptor proximal signaling reveals emergent ultrasensitivity. *PLoS Comput Biol* (2013) **9**(3):e1003004. doi:10.1371/journal.pcbi.1003004
67. Hogquist KA, Jameson SC, Heath WR, Howard JL, Bevan MJ, Carbone FR. T cell receptor antagonist peptides induce positive selection. *Cell* (1994) **76**(1):17–27. doi:10.1016/0092-8674(94)90169-4
68. El-Gogo S, Staib C, Meyr M, Erfle V, Sutter G, Adler H. Recombinant murine gammaherpesvirus 68 (MHV-68) as challenge virus to test efficacy of vaccination against chronic virus infections in the mouse model. *Vaccine* (2007) **25**(20):3934–45. doi:10.1016/j.vaccine.2007.02.054
69. Currie J. *Stochastic Modelling of TCR Binding*. Ph.D. thesis, Leeds: University of Leeds (2012).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2013; accepted: 15 March 2014; published online: 30 April 2014.

*Citation: Castro M, van Santen HM, Férez M, Alarcón B, Lythe G and Molina-París C (2014) Receptor pre-clustering and T cell responses: insights into molecular mechanisms. Front. Immunol. **5**:132. doi: 10.3389/fimmu.2014.00132*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Castro, van Santen, Férez, Alarcón, Lythe and Molina-París. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 04 February 2014
doi: 10.3389/fimmu.2014.00013

Theories and quantification of thymic selection

Andrew J. Yates*

Departments of Systems and Computational Biology, Microbiology and Immunology, Albert Einstein College of Medicine, New York, NY, USA

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Anton van der Merwe, University of Oxford, UK
Viktor Müller, Hungarian Academy of Sciences and Eötvös Loránd University, Hungary
Johannes Textor, Utrecht University, Netherlands

****Correspondence:***
Andrew J. Yates, Departments of Systems and Computational Biology, Microbiology and Immunology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, New York, NY 10461, USA
e-mail: andrew.yates@einstein.yu.edu

The peripheral T cell repertoire is sculpted from prototypic T cells in the thymus bearing randomly generated T cell receptors (TCR) and by a series of developmental and selection steps that remove cells that are unresponsive or overly reactive to self-peptide–MHC complexes. The challenge of understanding how the kinetics of T cell development and the statistics of the selection processes combine to provide a diverse but self-tolerant T cell repertoire has invited quantitative modeling approaches, which are reviewed here.

Keywords: thymic selection, T cells, mathematical modeling, repertoire selection, theoretical biology

INTRODUCTION

Conventional ($CD4^+$ and $CD8^+$) T cells are an integral part of adaptive immune systems in vertebrates. A key stage in their development is the creation of the T cell receptor (TCR) through a stochastic process of gene rearrangement. The resulting pre-selection TCR repertoire has the potential to recognize a very large array of peptides derived both from self and from foreign organisms, presented on Major Histocompatibility Complex (MHC) molecules on the surfaces of other cells. Much of T cell development occurs in a specialized organ in the chest called the thymus, within which this diverse potential repertoire of TCR is vetted. A process referred to as positive selection removes cells with TCR conformations that are generally non-responsive to self-peptide–MHC ligands (self-pMHC), and negative selection removes cells that are overly reactive to self-pMHC and pose a threat of autoimmune responses. The post-selection repertoire exported from the thymus comprises T cells that are largely non-responsive to self, yet capable of responding with remarkable specificity to foreign peptides.

There is a very extensive literature relating to thymic development and selection [for reviews, see for example Ref. (1–3)], but here we summarize the key ideas briefly (**Figure 1**). Conventional T cells begin life as lymphoid progenitors, which migrate from the bone marrow to the inner, cortical region of the thymus and begin a process of proliferation and maturation. Early in development in the cortex thymocytes are referred to as double negative (DN), lacking expression of the CD4 and CD8 co-receptors that are involved in TCR signaling. The TCR comprises two chains and is formed by a multi-step gene rearrangement process that first generates the TCRβ, γ, and δ chains (a small proportion of cells diverge at this stage to seed the $\gamma\delta$ T cell lineage) and then the TCRα chain at around the transition from the DN to $CD4^+CD8^+$ (double positive, DP) stage. TCR$\alpha\beta$ cells then migrate among cortical thymic epithelial cells and dendritic cells, auditioning for the ability to recognize self-pMHC. There is evidence that DP cells with non-functional TCR can undergo repeated TCRα rearrangements (4) to re-audition. Positively-selected cortical thymocytes begin negative selection and eventually move to the outer capsule of the thymus, the medulla. There they complete negative selection through interactions with medullary thymic epithelial cells and dendritic cells. TCR$\alpha\beta$ thymocytes, which recognize self-peptides presented on MHC class I or class II below an acceptable threshold of reactivity develop into the CD8 SP (single-positive, $CD4^-CD8^+$) or CD4 SP ($CD4^+CD8^-$) lineages, respectively, and are eventually exported into the peripheral circulation as naive T cells.

The topic of thymic selection has received substantial attention from the immunological modeling community, perhaps for two main reasons. First, selection has widely been viewed as a well-delineated optimization problem – how to craft a TCR repertoire that covers the space of possible pMHC ligands as widely as possible, while preserving sufficient specificity to discriminate between self and foreign (and between different foreign) peptides? This question naturally invites quantitative arguments. Second, the biology is well-characterized – a relatively small number of cell types and modes of interaction appear to be involved, and large amounts of experimental data are available. These simplify and constrain the construction of models.

Modeling studies have focused on many aspects of thymic selection but many questions and uncertainties remain. What are the rates and efficiencies of passage through the different phases of development and selection, and in what thymic microenvironments do each take place? How do thymocytes integrate signals received from interactions with pMHC to make fate decisions?

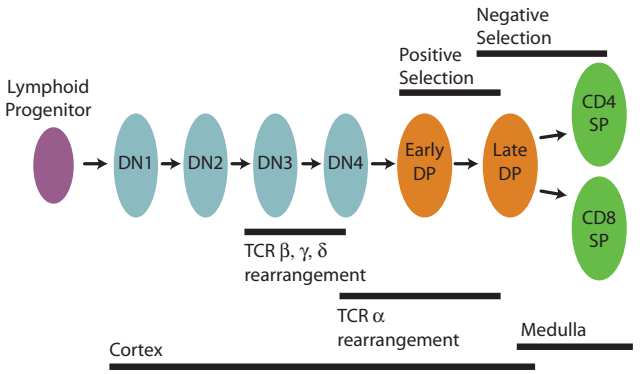


FIGURE 1 | Stages in the development of CD4 and CD8 T cells in the thymus.

What are the relative contributions of the MHC itself and its associated peptide to TCR signaling and fate determination? What influence do each of these have on the post-selection repertoire's diversity and coverage of the pMHC universe, and its ability to discriminate between self and foreign? How complete is the removal of potentially self-reactive clones? How many TCR interactions contribute to a thymocyte's fate decisions? What evolutionary pressures have determined the typical number of MHC alleles we possess? There have been many different theoretical approaches to these questions – from mean-field population dynamic models of progression through developmental stages, to probabilistic models of selection, to explicitly spatial models of migration within the thymus.

This review groups studies of these topics into broadly labeled categories, but in some cases the grouping is arbitrary – many of these questions are related and have been addressed either alone or in combination. The review has a bottom-up structure, beginning with an overview of experimental quantification of selection and modeling of thymocyte population dynamics. It then moves to studies of higher-level properties of the T cell repertoire, such as TCR cross-reactivity, and concludes with the problem of optimal within-individual MHC diversity.

THE POPULATION DYNAMICS OF THYMOCYTES

Basic elements of a quantitative understanding of thymic development are the steady-state population sizes of different developmental stages, the mean times to transit between them and the proportion surviving at each stage, which we refer to as the efficiencies of selection. While some quantities can be experimentally determined, mathematical models have helped us develop a more complete description of the kinetics of selection, both for the thymocyte population as a whole and for the CD4 and CD8 lineages in isolation.

To estimate the parameters of a dynamical system usually involves observing its response to perturbations. One method is to follow cohorts of cells as they progress through development using intra-thymic injection of a dye or radioisotope label (5–8). Arguably this method is less disruptive than cell transfers, but the uptake of marker can be heterogeneous (5, 7) and measurements of death rates using injected dyes rather than congenic markers may be confounded by loss of label (9). More recently, methods have included using GFP (green fluorescent protein) expressed during TCR rearrangement, its decaying intensity then a marker for time spent in development (10); inducible TCR signaling can be used to arrest, release, and follow cohorts of cells from the early DP stage (11); and small numbers of labeled thymocytes isolated at different developmental states can be followed after intra-thymic injection (11, 12). The population dynamics have also been exposed by transiently depleting thymocytes and observing the system's return to equilibrium (13). Various experimental systems, with or without associated dynamical models, are in general agreement over several quantitative aspects of thymic development but inconsistencies and uncertainties remain.

SELECTION EFFICIENCIES AND CELL FLUXES

Thymocytes begin to select against self-pMHC ligands at the DP stage following TCR rearrangement and so we focus on survival, proliferation, and differentiation from this stage onward. The proportion of DP cells that reach maturity (that is, survive both positive and negative selection) is widely agreed to be 5% or less (6, 11, 13–16). Within this pruning process, the general view is that positive selection is the most stringent, with 75–80% of cells failing to progress from the earliest DP stage, suggesting the majority of TCR generated are unable to recognize peptides in conjunction with MHC class I or II to any useful degree (11, 13, 15, 17, 18). Many studies have estimated that between 20 and 50% of positively-selected thymocytes then survive negative selection (11, 17, 19–22), although Itano and Robey (8) estimated a selection efficiency as high as 90% for DP cells into the CD4 SP lineage.

The rate of production of mature CD4 and CD8 cells in the thymi of young adult mice is roughly 1% of total thymocytes or $1-3\times10^6$ cells/day, a figure arrived at by a variety of labeling methods (5–7). Egerton et al. (6) estimated this to be just over 3% of the rate of entry into the DP population, meaning that fueling this trickle of output requires that roughly 30% of all thymocytes enter the DP stage each day. This again illustrates the extent of the filtering of the pre-selection repertoire that appears to be required to produce a functional and self-tolerant population of naive T cells. The thymus gradually involutes and its rate of output declines with age in both mice (23) and in humans (24), indicating that the bulk of the peripheral T cell repertoire is probably generated early in life.

THE MAJORITY OF THYMOCYTE DIVISION LIKELY OCCURS PRE-SELECTION

Labeled nucleotide uptake assays have revealed that substantial proliferation of thymocytes occurs before selection on self-pMHC ligands begins, stopping at or around the time of TCR rearrangement at the late DN/early DP stage (6, 25, 26). However, it is proliferation following TCR rearrangement that is most relevant for understanding how repertoire diversity is generated. Division during selection means a smaller proportion of TCR clonotypes may pass selection than measures of percentage survival suggest (27). The extent of division early in selection is unclear – estimates of the proportion of newly generated DP cells that are dividing have ranged from 11 to 68% (6, 25, 28), and CFSE labeling in

in vitro thymic organ cultures showed up to 5 divisions from DP onward (29). However, the DP population comprises cells pre- and post-TCR rearrangement, and there appears to be very little proliferation within the more mature DP population (6, 11, 15). There is a low level of proliferation during or just before the SP stage (11, 13, 25, 30), with CD8 SP more prone to division than CD4 SP (10).

Perhaps the most reliable experimental measure of average levels of proliferation during selection uses T cell receptor excision circles (TRECs). TRECs are circular DNA fragments that are stable remnants of the recombination events that generate the TCR and are shared randomly between daughter cells on division. The mean TREC content per cell is a rough measure of the mean number of divisions that have taken place since the TCR was generated. One caveat is that TREC studies are used most commonly in humans and much of what we discuss here derive from studies in mice. Another is that standard TREC measurements contain no information about the variance of the division number, and may gloss over even quite extreme heterogeneity in division patterns. Nevertheless a study of human infants observed 1–2 divisions on average between TCR rearrangement at the $CD3^{low}$ $CD4^+CD8^+$ stage and mature CD4 or CD8 SP; once shortly after TCR rearrangement, and another at the CD8 (but not CD4) SP stage (31). The high TREC content they observed at the early DP stage may reflect multiple rearrangements taking place in order to generate a functional TCRα-chain. In line with these results, the TREC content of naive $CD31^+CD4^+$ recent thymic emigrants in human infants is ~0.1–0.9/cell (32), suggesting that up to three divisions take place on average between TCR rearrangement and export to the periphery, although this may include some post-thymic proliferation and so is an upper limit on the extent of intra-thymic division.

TURNOVER RATES AND TRANSIT TIMES

Experimental estimates of the times taken to transit different developmental stages (immature DP → mature DP → SP → Export) are variable, particularly within the SP population (6, 10, 12, 25). Possible reasons for these discrepancies include different labeling protocols, different gating strategies defining thymic subpopulations, heterogeneity of cell populations, and differences in the kinetics of MHC class I-restricted and class II-restricted lineages. It has also been unclear whether selection is a "conveyor belt," first-in first-out, or has a more stochastic "lucky dip" nature (25). From a modeling perspective these are two points on a continuum. If an experimentally identifiable developmental stage comprises several shorter, sequential differentiation steps, the variance in the transit time through that stage is lowered with respect to a single-step model of transit. The more obligate steps, the more conveyor-belt-like the system appears.

There is general agreement that the transition from non-dividing mature DP to SP takes on average 3–4 days (6, 12, 15, 28), although it has been argued that it takes significantly longer to reach CD8 SP than CD4 SP (33). This transition is dependent on TCR signaling (15, 34). Observing a well-defined delay in the appearance of labeled SP cells, Egerton et al. (6) argued for a first-in-first-out kinetic in the DP population. This suggests DP cells must transit through a number of obligate steps. Subsequent experimental and modeling studies have addressed this, and are discussed below. The same study estimated a mean SP residence time of ~12 days, comparable to other estimates of the medullary residence time (6, 28). McCaughtry et al. (10) argued that this is an overestimate of the time mature conventional SP T cells take to develop, because the SP population is heterogeneous, also containing Treg, NKT, and $\gamma\delta$ T cells, which turn over more slowly. They estimated SP CD4/CD8 residence times to be 4.4/4.6 days. Saini et al. (33) arrived at similar estimates. As for DP cells, there are may be several developmental stages within the SP population and so it seems unlikely that SP residence times are exponentially distributed.

Stritesky et al. (12) estimated the total rate (cells per unit time) at which cells are negatively selected to be almost six times greater than the rate of positive selection, and found that both processes occur predominantly at the DP stage. Converting these figures into the relative efficiencies of positive and negative selection requires knowledge of how long cells spend in each selecting phase. If indeed positive selection is the more stringent, their result indicates that negative selection must take place over a relatively short timescale within the DP compartment. This is supported by a recent study observing negative selection of DP thymocytes taking place over ~12 h (35).

Interpreting data on transit or residence times can be problematic when both death and differentiation are taking place, as they clearly are at the DP stage(s) of development. If death and differentiation are modeled as independent processes, then at equilibrium transit rates through a compartment are not necessarily the same as turnover rates. If cells are maturing at rate μ and dying at rate δ, the population turns over at rate $\mu + \delta$ and the expected time a cell spends in that compartment is $1/(\mu + \delta)$. However, the mean time that successfully differentiating cells spend in each compartment is shorter because it is conditioned on survival, and is $\mu/(\mu + \delta)^2$ (if cells are capable of maturing but are simultaneously at risk of dying, those that successfully mature tend to do so early). This difference can be quite substantial, as we see below.

KINETIC MODELS OF THYMIC DEVELOPMENT

Data from these experimental studies and others have invited the use of population dynamic models to infer the kinetics of development. In the first studies to model thymic development, Mehr and collaborators utilized ordinary differential equation (ODE) models of the flow from DN → early DP → late DP → CD4/CD8 SP (36, 37). They utilized measures of steady-state population sizes and parameters either inferred from data or explored systematically to ask questions about the underlying dynamics. Mehr et al. (36) argued that positive selection likely involves triggering of proliferation as well as rescue from death, and while they were unable to use the steady-state data to make strong statements about the timing of positive versus negative selection, they inferred that most death at the DP stage is due to failure to positively select, consistent with many experimental and subsequent modeling studies.

There is evidence from fetal thymic organ cultures that populations of mature $CD4^+$ T cells resident in the thymus may enrich for the CD4 lineage while reducing thymic output. Mehr et al. (37) used a similar model with these data to propose that the mature resident cells increase survival of developing single-positive CD4 T cells while reducing proliferation or increasing the rate of

differentiation of DP cells. They suggest that mature CD4 T cells exert their influence by restricting the number of available pMHC ligands in the thymus, which could simultaneously reduce proliferation of DP cells (lowering thymic output) and decrease the stringency of negative selection (increasing the efficiency of maturation into the mature SP state). Again, these conclusions were reached using data from the thymus at steady-state.

Mehr and collaborators also studied the seeding of the cortical stroma with bone marrow-derived progenitor cells using a combination of modeling and experiment. They showed how migration between niche sites explained the competitive advantage of younger progenitors over older (38, 39), and that reconstitution of the progenitor population following irradiation is limited by damage to stromal niches and incumbent, surviving cells (40).

Thomas-Vaslin et al. (13) studied naive T cell homeostasis from the thymus through to the periphery. They induced systemic depletion of T cells for 7 days through expression of a suicide gene in dividing cells, and followed the kinetics of reconstitution. To interpret these data they developed a multi-compartment ODE model of T cell development, with a finer-grained treatment of transit through the DN, DP, and SP stages. In their model extensive proliferation occurs through the DN to early DP, with the latter population dividing 5 times. Their best-fitting model assumes all cell death (positive and negative selection) takes place at the late DP stage. They estimated 5% of total thymocytes (DN, DP, and SP) or $\sim 3 \times 10^6$ are exported as naive SP cells per day, and that 93% of DP thymocytes are lost, in line with existing estimates, and again suggesting that the bulk of negative selection occurs at DP. The mean times spent overall in the early DP (dividing), late DP (selecting), and SP compartments were estimated to be 1.2, 2.7, and 5.8 days respectively.

Sinclair et al. (11) used a different experimental system, with controllable TCR signaling that allowed arrest and release of cells at the early DP stage, and used a multi-compartment ODE model to quantify transit dynamics and selection efficiencies. Rather than simply early or late, they broke the DP stage into a branched developmental progression defined by the expression levels of CD5 and the TCR (33). In their schema, DP1 thymocytes are pre-selection; progression to DP2 requires a positively-selecting TCR signal; DP2 thymocytes consist of class I- and class II-restricted thymocytes in the first 12–48 h of development; and DP3 thymocytes are predominantly MHC class I-restricted cells that can select into CD8SP only. Thus cells destined for CD4SP transit DP1-DP2 only, and CD8SP transit through DP1, DP2, and DP3.

Sinclair et al. (11) estimated that ~75% in DP1 fail to progress to DP2, reflecting failure to positively select and dying of neglect. Overall, 5% of DP cells become CD4SP and ~2% become CD8SP, and so ~94% of DP cells are lost. They also saw relatively low levels of cell death in the SP compartment. These results suggest again that the bulk of negative selection occurs before cells transition to SP. They saw very little proliferation in their system, using a variety of methods, and so did not model cell division. Mean residence times in DP1 and DP2 were 3.5 and 1.4 days, respectively, with the smaller CD8 lineage spending an additional 7 days in DP3. They estimated 23% of all thymocytes at DP and SP enter the DP compartment per day. These selection efficiencies and the net flux agree with other estimates. Accounting for the selection bias on maturing cells, the model predicts that successful thymocytes spend on average 1.3 days in DP1 + DP2, 4.5 days in DP3. SP4 and SP8 residence times were 5 and 3.7 days, respectively, with very little cell death occurring. Their analysis therefore suggests that CD4SP/CD8SP cells take ~6.3/9.5 days from entry into DP1 to export.

MIGRATION WITHIN THE THYMUS AND THE TIMING OF POSITIVE AND NEGATIVE SELECTION

From the perspective of modelers attempting to connect models of thymocyte dynamics to data, it is important to understand when and where the different phases of development and selection occur. Selection begins in the thymic cortex, where the majority of thymocytes perform undirected random walks (41) encountering pMHC on cortical thymic epithelial cells. Sensitivity to medullary chemokine receptor signals begins to increase immediately following receipt of a positive selection signal and positively-selected cortical thymocytes eventually display rapid, directed motion toward the medulla (41), where they encounter pMHC on medullary thymic epithelial cells and dendritic cells. Negative selection takes place in the medulla (35, 42–44) but also late in migration through the cortex (45) and possibly even throughout development (46). The mapping between these migratory and selecting processes to developmental stages is not clearly defined. Cells undergoing negative selection in the medulla include DP populations (35), indicating that maturation from DP to SP does not coincide precisely with the cortical–medullary transition but further supporting the conclusion that the extensive cell loss at the DP stage comes from failure of both positive and negative selection. Further, antigen-presenting cells in the cortex and medulla appear to differ in their ability to provide positive or negative selection signals, either through differences in pMHC expression or diversity, or levels of co-stimulation (47–51). It seems therefore that negative selection at the DP stage takes place in at least two distinct spatial and TCR-stimulatory environments.

MODELS OF SELECTION WITHIN THE CORTEX AND MEDULLA

Motivated by this, Faro et al. (52) took a different perspective; rather than partitioning selecting thymocytes into developmental stages, they used a probabilistic model to describe selection within the cortex and the medulla. They aimed to quantify the number of selecting events, the number of selecting APC encounters and pMHC engagements, and the efficiencies of positive and negative selection in each region. Using the experimental estimates of overall selection efficiencies, and one experimental estimate of the efficiency of negative selection in the medulla, they inferred that most thymocyte death occurs by failure to positive select in the cortex, and cells are ~10 times more likely to be deleted (negatively selected) in the medulla than in the cortex. With these efficiencies, through a parameter search, they were able to infer the number of ligands each thymocyte selects on in each spatial compartment. They came to the striking conclusion that for each cortical thymocyte selection takes places on <60 pMHC ligand interactions, likely in order to achieve in their model the required high level of failure to positively select. However, this needs to be reconciled with the ~3-day mean lifetime of cells at DP1, which suggests cells have far more opportunities to positively

select, either through repeated encounters with APC or through repeated rearrangements of the TCRα chain [see Ref. (53) and refs therein], before dying of neglect.

IDENTIFYING THE SOURCE OF THE CD4:CD8 LINEAGE BIAS IN THYMUS

CD4 SP outnumber CD8 SP by roughly 4:1 in the thymi of many species. Using time courses of development in control mice and those lacking MHC class I or class II, Sinclair et al. (11) estimated the CD4 and CD8 lineage-specific selection efficiencies. In control animals, the highest death rate was at the positively-selected DP2 stage, and was substantially greater for MHC class I-restricted cells. MHC class I- and class II-restricted cells are indistinguishable at DP1 and DP2, but they were able to back-calculate the rates of production of precursors of the two lineages after TCR rearrangement, and found they were comparable. This suggests that the CD4:CD8 asymmetry in the thymus derives in large part from more stringent selection acting on MHC class I-restricted cells and not from any significant asymmetry in the predisposition of randomly generated TCR to recognize MHC class I or class II. Theirs is a model of CD4/CD8 lineage commitment in which the ability of a DP thymocyte to recognize MHC class I or class II dictates whether it will progress to the CD8 or CD4 lineages, respectively (8, 54). This is contrast to a less efficient, selective process in which a thymocyte's decision to downregulate either CD4 or CD8 expression is stochastic and decoupled from MHC preference, such that potentially viable TCR may fail positive selection [see, for example Ref. (55, 56); and Ref. (57) for a discussion of a hybrid mechanism]. Mehr et al. (36) proposed a purely instructive model of selection, in which pre-selection thymocytes are in principle able to recognize both MHC class I or II, and concluded that the most likely explanation of the CD4 bias is a difference in the *per capita* rates of maturation from DP into the two lineages, rather than differences in death rates.

The majority of models discussed here assume that thymocytes undergo screening independently. Mehr et al. (36, 37) implicitly allowed for competition with density-dependent proliferation rates at each developmental stage. However, there is some evidence that the probabilities of maturation can be impacted by competition between thymocytes, both globally and in lineage-specific ways. The efficiency of selection of transgenic TCRs varies with their abundance and with the availability of cognate pMHC (15, 58–60), and the selection of polyclonal MHC class I-restricted thymocytes is more efficient in the absence of MHC class II and vice versa (11). These observations suggest that selection efficiencies may be limited by competition both within and between lineages for access to pMHC or other resources needed for selection, and so may impact on the CD4:CD8 ratio emerging from the thymus. Two studies have used explicitly spatial, agent-based models of thymocyte migration and development to investigate this issue. Souza-e Silva et al. (61) modeled the movement of DN, DP, and CD4 SP and CD8 SP populations and their interactions with thymic epithelial cells (TEC) and chemokine gradients, using a 2D model. The structure of the epithelial networks was derived from histological samples from both mice and infant humans. Parameters were chosen to give agreement with published data regarding the repopulation of the thymus after sublethal irradiation, although a sensitivity analysis was not performed. In their model the CD4:CD8 ratio emerges as a result of competition for access to TEC and stochastic variation in the duration of signaling, which has been associated with CD4/CD8 lineage commitment (62). Their simulations also reproduce an observed variation in the CD4:CD8 ratio as irradiated thymi reconstitute and, in their model, the degree of competition increases. Efroni et al. (63) also took an agent-based approach and concluded that MHC class I and class II ligands on TECs are limiting. If continued access to pMHC stimulation is required for survival, and class I restricted cells stay conjugated to MHC for longer than MHC class II-restricted cells, exclusion of competitors leads to a higher death rate of cells developing into the CD8 lineage and a skewing of the CD4:CD8 ratio. Such a competitive model is an experimentally testable explanation of the differential death rates observed by Sinclair et al. (11).

CHARACTERISTICS OF THE TCR REPERTOIRE

Various summary statistics can be used to describe T cell populations pre- or post-selection. The *diversity* (or the *repertoire*) usually denotes the total number of distinct TCR sequences or clonotypes. The *cross-reactivity* measures a TCR's capacity for discrimination, and is quoted as either the average number or the proportion of different pMHC that one TCR responds to above some defined functional threshold. *Specificity* is inversely related to cross-reactivity. A mirror quantity is the *precursor frequency*, also referred to as the response frequency – the average proportion of all TCR capable of recognizing one pMHC. Further, selection operates in the context of an individual's own MHC alleles. *MHC restriction* measures the degree to which a given TCR is limited to recognizing peptides presented by one or more self-MHC; and *alloreactivity* is the proportion of TCR that respond to a foreign MHC, which is relevant for transplantation of tissues from one individual to another. In the sections that follow we describe how theoretical models have been used to understand how these quantities are linked and constrained by thymic selection.

TCR CROSS-REACTIVITY

A diverse TCR repertoire seems to be a requirement for coverage of pMHC shape space. However, the number of theoretically possible pMHC complexes appears to be far greater than any individual's capacity for unique TCR clonotypes (64–66); a simple calculation for just one MHC class I variant, assuming it presents 2% of all possible 9-residue peptides, yields $20^9 \times 0.02 \simeq 10^{10}$ possible pMHC, compared with the roughly 5×10^7 naive CD8 T cells in a mouse. To minimize the probability that any given foreign pMHC will escape detection by the immune system, some degree of TCR cross-reactivity therefore seems beneficial. Mason (64) used a variety of methods and data sources to estimate that one MHC class I-restricted T cell responds to between 10^6 and 10^7 nonamer peptides, or one in 10^3 to 10^4 pMHC using the theoretical estimate of the potential pMHC diversity; and Ishizuka et al. (65) used peptide libraries to estimate more directly that one CD8 T cell clone responds to roughly 1 in 3×10^4 peptide–MHC class I ligands. On the other hand, the average degree of cross-reactivity seems necessarily constrained from above, to avoid excessive deletion of the repertoire and to preserve specificity for self/non-self discrimination. It therefore seems plausible that evolutionary pressures

might have optimized this trade-off and determined the degree to which TCR can respond to multiple pMHC.

OPTIMAL LEVELS OF TCR CROSS-REACTIVITY – PROBABILISTIC ARGUMENTS

Several variants of essentially the same argument predict that the diversity of self-peptides involved in selection is the strongest influence on the optimum level of TCR cross-reactivity (64, 67–70). One version of the argument is as follows. The proportion of the positively-selected T cell repertoire R_0 that avoids deletion, f, decreases with both the number of self antigens N_s and the cross-reactivity r, $f = (1-r)^{N_s}$ which is approximately $\exp(-rN_s)$ for $r \lesssim 1/N_s$. A pathogen escapes immune recognition if all fR_0 surviving unique clonotypes fail to recognize (cross-react with) all x epitopes it generates, with probability

$$P_E = (1-r)^{fR_0 x} \simeq \exp(-rfR_0 x) \tag{1}$$

where again the approximation holds if $r \lesssim 1/(fR_0 x)$. This ignores MHC restriction, but including this refinement yields similar conclusions (67). Using the expression for f,

$$R_0 \simeq -\log(P_E)\frac{\exp(rN_s)}{rx}. \tag{2}$$

This equation connects the repertoire before negative selection R_0, the probability of immune escape P_E and the pre-selection cross-reactivity r. R_0 is relatively insensitive to P_E but very sensitive to the diversity of self, N_s. In this model, then, the strongest determinant of the size of the pre-selection repertoire is the diversity of self antigens, N_s, and not the requirement for minimizing the probability that a pathogen escapes detection (67).

The three-way relation expressed by equation (2) can then be used to estimate the optimal cross-reactivity under different evolutionary constraints. Suppose the potential repertoire size R_0 is relatively conserved and evolution has selected for the smallest P_E by tuning TCR cross-reactivity; in this case, the optimal cross-reactivity is simply the inverse of the number of distinct self-pMHC involved in selection, $r = 1/N_s$. The same value of r arises if evolution is assumed to minimize the required repertoire size R_0, whatever the value of P_E (67). Thus the more diverse the self-peptides involved in thymic selection, the more specific (less cross-reactive) the TCR needs to be. The same result can be derived in a very general way using extreme-value theory (70), requiring only the assumption that the negative selection threshold in the thymus is equal to the activation threshold in the periphery.

The induction of tolerance in the thymus is likely incomplete and there may be mature lymphocytes that are able to recognize self-peptides not involved in thymic selection. Borghans and De Boer (71) argued that to minimize the probability of these cells mounting a cross-reactive autoimmune response to this "ignored self" while responding to a pathogen demands higher levels of specificity than predicted by the simplest models. In this model, optimal cross-reactivity is then modulated by the potential diversity of the repertoire; the greater the number of possible T cell clonotypes, the lower cross-reactivity is required.

Percus et al. (72) took a different approach to studying optimal cross-reactivity, prompted by the observation that the sizes of the binding sites of the TCR and the B cell receptor (antibodies) are similar, at roughly 15 amino acids. They concluded that this size is large enough to provide discriminatory power but small enough that there is sufficient cross-reactivity for coverage of foreign antigen shape space. Interestingly this result does not arise from the demand for self–non-self discrimination, but rather from the constraint of the observation that the B and T cell repertoires comprise ~10^7 different receptors. However, this diversity itself may be derived from the self-tolerance arguments described above (64, 67–69). It has since been established that substantially fewer peptide residues are involved in TCR recognition. Burroughs et al. (73) analyzed the proteomes of humans and several microorganisms and showed that even the seven exposed (non-anchor) residues of the nine-mer peptides bound to one MHC class I allele may promote self/non-self discrimination, with <0.5% overlap in these sequences between humans and different microorganisms.

CONVERGENT ESTIMATES OF LEVELS OF NEGATIVE SELECTION

Several of these studies concluded that at the optimal level of cross-reactivity the probability of negative selection is roughly 63%, making various assumptions regarding the magnitude of parameters and maximizing the probability that the post-selection repertoire mounts a response to a foreign pMHC. However, the probability of negative selection can be derived without any assumptions regarding parameter values. From above, the fraction of the positively-selected repertoire with cross-reactivity r that survives deletion on N_s self-peptides is $f = (1-r)^{N_s}$. The probability that the post-selection repertoire $R = fR_0$ fails to recognize one given foreign pMHC is given by equation (1) with $x = 1$,

$$P_E = (1-r)^{fR_0} = (1-r)^{R_0(1-r)^{N_s}}. \tag{3}$$

This is minimized with respect to r at $r = 1 - \exp(-1/N_s)$, exactly (the optimal cross-reactivity $r \simeq 1/N_s$ then obtains if $N_s \gg 1$). So if evolution acts on cross-reactivity to minimize the probability of foreign pMHC escaping detection, the fraction of the positively-selected repertoire that survives negative selection is then simply $f = (1-r)^{N_s} = \exp(-1) \simeq 0.37$, or $\simeq 63\%$ of positively-selected thymocytes are deleted.

Mason (64) arrived at the same result assuming heuristically that the quantity to be maximized is the "reactivity" of the repertoire, proportional to the number of peptides each T cell can recognize multiplied by the proportion surviving negative selection;

$$\text{Reactivity} \sim \text{Cross-reactivity} \times \text{P(survive negative selection)} \sim r \times (1-r)^{N_s}.$$

Maximizing this reactivity is equivalent to minimizing the probability of escape in equation (3) when r is assumed to be small. There, using the Taylor expansion gives $P_E \simeq 1 - rR_0(1-r)^{N_s}$, and so the probability of responding $(1 - P_E)$ is $\sim rR_0(1-r)^{N_s}$, or Mason's reactivity. Since r is small, the probability of negative selection is $(1-r)^{N_s} \simeq \exp(-rN_s)$ and so the reactivity is proportional to $r\exp(-rN_s)$, which is maximal with respect to r when argument of the exponential is -1. Thus again $f \simeq 0.37$ and the optimal cross-reactivity $r \simeq 1/N_s$.

An essentially identical argument applies to negative selection of B cells (67, 69). This estimate of f is remarkably consistent with estimates of levels of negative selection in the thymus from several experimental and population dynamic modeling studies (11, 17, 19–22).

ALTERNATIVE TREATMENTS OF CROSS-REACTIVITY

These models assume a universal cross-reactivity parameter r, but T cells may have the capacity to modulate their activation thresholds in response to their signaling environment (74, 75). Motivated by this, Scherer et al. (76) developed a model in which T cells tune their activation thresholds (and thus their cross-reactivity) to the level of their strongest interaction with self-pMHC during selection. If combined with a deletion mechanism that removes cells with activation thresholds so high as to be judged functionally inert, this model appears to be a more efficient mechanism of thymic selection than the standard clonal deletion model. Scherer et al. showed that the tuning model increases the probability of mounting an immune response to a given pathogen epitope, given a pre-selection repertoire size R_0, and the number of self-pMHC ligands involved in selection, N_s. The improvement offered by the tuning model is most striking for small pre-selection repertoires, $R_0 \ll N_s$, but disappears for $R_0 \gg N_s$. The latter inequality likely holds for mice and humans; the potential number of unique TCR sequences exceeds the estimated 10^3–10^5 self-peptides able to be presented by a given MHC allele (73, 77, 78). Further, equation (1) predicts that at the optimal cross-reactivity $r = 1/N_s$, the probability of one epitope ($x = 1$) escaping recognition is $P_E = \exp(-R_0/eN_s)$ where e is the base of the natural logarithm. For $P_E < 0.05$, expected in humans and mice, requires $R_0 \gtrsim 10N_s$. Despite this, Scherer et al. (76) argue that the tuning model is a more parsimonious mechanism of self-tolerance in the thymus than the standard model of deletion based on evolutionarily optimized cross-reactivity.

Finally, many of these arguments assumed thymic selection alone optimizes cross-reactivity, but the requirement for memory T cells to discriminate between different pathogens may impose a further constraint of its own (79, 80).

EXPLORING CROSS-REACTIVITY WITH SEQUENCE-BASED MODELS OF THYMIC SELECTION

A series of related papers by Detours, Perelson, and Mehr (27, 81–83) used a model of TCR–pMHC interactions to understand at a more mechanistic level how cross-reactivity, alloreactivity, and MHC restriction emerge in the post-selection repertoire. Here we focus on their treatment of TCR cross-reactivity, and return to alloreactivity and MHC restriction in the next section. Their starting point was an established model of protein binding (81, 84). They described the interaction between the variable region of the TCR and its pMHC ligand with strings of digits, and binding strengths between each digit pair were determined by the degree of complementarity between their binary representations (81). MHC and peptide contributed additively to the affinity of the interaction, the quantity assumed to drive selection. Given the number of digits ascribed to the polymorphic MHC residues in contact with the TCR, and the number of digits representing the peptide, selection could be performed on a randomly generated TCR repertoire using randomly generated peptide–MHC complexes. Affinity thresholds were then adjusted to give stringencies of positive and negative selection similar to those observed experimentally.

To circumvent the computational costs of selection using realistic numbers of peptides and unique pre-selection TCRs, they derived expressions for the mean-field predictions of the model for given parameter sets. This has the advantage of yielding population-level statements, which average over all possible TCR, MHC, and peptide sequences.

Detours and Perelson (82) estimated the precursor frequency, the proportion of naive T cells able to respond to a particular foreign pMHC. Experimental estimates of this quantity lie in the range 10^{-6}–10^{-4} (85–89). They term this the response frequency, R, and found it to be strongly and inversely related to the number of selecting self-pMHC ligands. Since precursor frequency is positively correlated with cross-reactivity (64), this result is in keeping with the theoretical studies discussed above (64, 67–70). It is also consistent with observations that repertoires selected on a restricted range of peptides exhibit higher cross-reactivity than normal (90–92). For R to lie in the observed range constrains the number of distinct peptides each MHC can present to be of the order 10^3–10^5, in line with estimates for murine MHC class I (77), MHC class II (78), and human MHC class I (73).

To explore the effect of thymic selection on specificity in more detail, Chao et al. (93) revisited the complementary digit-string model. Again peptide and MHC were assumed to contribute additively to an antigenic distance from the TCR, which was inversely related to affinity or the strength of a selecting signal. They confirmed that negative selection reduced the coverage of peptide space, defined as the proportion of peptides that are recognized on the selecting MHC. This was equivalent to a reduction in the cross-reactivity of the repertoire; it reduced the mean antigenic distance to foreign pMHC complexes.

Chao et al. (93) then used the model to address the question of why the number of pMHC that one T cell is able to respond to varies widely across TCR (94). Their simulations suggested that the degree of cross-reactivity to a foreign peptide was inversely related to the peptide's similarity to self, which can be understood with the following argument. In their model, in the pre-selection repertoire a TCR's affinity for the MHC and peptide portions of its ligand are uncorrelated. Selection introduces an inverse correlation between a TCR's affinity for its selecting MHC and its strongest affinity for self-peptide; to select, a TCR's strongest interaction with self must lie between the positive and negative selecting thresholds. (The narrower the range of affinities defining the selecting region, the stronger this correlation will be.) Selected T cells with high affinity for MHC then have a relatively low affinity for the self-peptide component and require only weak binding to foreign peptide to be activated (activation in their model is defined to be an interaction above the negative selection threshold). These cells are therefore cross-reactive to foreign peptides. Conversely, selected TCR that bind relatively weakly to MHC have higher affinity to self and require strong binding to foreign peptide for activation, and therefore have more specificity for foreign antigen. Thus it emerges from their model that a TCR's specificity to foreign peptide is positively correlated to its affinity for self-peptide; or

equivalently, a TCR's cross-reactivity is positively correlated with its affinity for MHC.

The effect of negative selection on cross-reactivity can be understood with a similar argument. A TCR with high affinity for MHC will survive negative selection only if it has low affinity to all self-peptides, which is unlikely. Negative selection therefore enriches for cells with lower affinity for MHC, which from the argument above tend to be less cross-reactive. This reduction in coverage means specificity to foreign peptide must be increased.

Kosmrlj et al. (95) used a more physical, mechanistic approach to understanding how negative selection increases specificity, with the aim of characterizing the properties of the amino acid sequences of specific and cross-reactive TCR. Using the Miyazawa–Jernigan matrix (96) to quantify the interaction energies of pairs of amino acids, they extended the digit-string model to calculate the binding affinities between the peptide and the CDR3 region of the TCR, with a constant contribution from the MHC. (The variable peptide element of the pMHC ligand can be assumed to include the polymorphic MHC residues; thus their model may allow for MHC restriction, although this was not discussed.) Košmrlj et al. (97) presents an analytical treatment of the model.

They observed that TCRs selected against multiple peptides on the same MHC had peptide contact residues enriched in weakly interacting amino acids. In their model this arises by a sort of buffering mechanism – such sequences are able to withstand multiple substitutions in the peptide sequence to which they bind most strongly, and so are more resistant to negative selection than those TCR with strongly binding residues. For these TCR to survive selection requires that the invariant MHC contribution to the binding energy is of moderate strength – contributing sufficiently to favor positive selection but well below the negative selection threshold.

Kosmrlj et al. (95) argue that it is this enrichment for weakly binding TCR driven by negative selection that underlies antigen specificity. Antigen recognition is assumed to occur when a TCR signal exceeds the negative selection threshold made up by several interactions. This requires the peptide to contain several amino acids capable of binding the most strongly to the generally weakly binding TCR contact residues. Each contributes significantly to the total binding energy, and so any mutation to the peptide sequence has a high probability of abrogating recognition. Thus there is a restricted peptide signature or "barcode" required to trigger the TCR. In their model, TCR selected against a single pMHC were enriched slightly for strongly interacting amino acids. For these TCR, they argue, fewer amino acids contribute on average to the binding energy, triggering is more robust to mutations in the peptide sequence, and so the TCR is more cross-reactive. Thus again the argument emerges that the cross-reactivity is inversely related to the diversity of self driving selection. Kosmrlj et al. (98) employed this idea to put forward an explanation of why the population of elite-controllers of HIV infection is enriched for the HLA-B57 allele. Using a predictive peptide binding algorithm they argued that HLA-B*5701 binds a lower diversity of self-peptides than average. Cytotoxic T cells restricted to this allele are then expected to be more cross-reactive than average and so are more resistant to virus mutations that might otherwise escape CTL control.

Chao et al. (93) and Kosmrlj et al. (95) took different approaches to the problem of how negative selection increases specificity. They came to the common conclusion that the most specific TCR are those with low to intermediate affinity to MHC – high enough to have a reasonable probability of passing positive selection, but low enough to avoid negative selection by allowing headroom for the additional contribution from the peptide component. The greater this headroom, the smaller the proportion of peptides that can trigger activation and so the greater the specificity.

THE EMERGENCE OF SPECIFICITY IN AVIDITY-BASED MODELS OF SELECTION

Van den Berg et al. (99) developed a statistical framework to study the question of how specificity and self-tolerance can derive from a pre-selection repertoire of relatively promiscuous TCR. In their formalism, T cell activation is avidity-based and related to the rate of TCR triggering. Their starting point is that TCRs are degenerate and low affinity, binding weakly to many pMHC. TCR perceive an average signal derived from endogenous self-pMHC, and are triggered only by pMHC with sufficiently high prevalence and affinity to be visible above this background. The authors introduce the concept of an antigen presentation profile (APP), characterizing the abundances of different pMHC on antigen-presenting cells (APC). Positively-selected cells are selected against a given number of APC each with distinct APPs. In their framework, negative selection acts only on ubiquitous peptides presented on all APCs, and decisions are made on the basis of the entire APP of one APC. TCR that are triggered by this constitutive self-background are deleted. This filtering acts to sharpen the boundary between triggering rates, which give low and high activation probabilities, and so specificity can emerge even from a highly degenerate TCR. Interestingly they predict that negative selection does not have to be particularly stringent to generate an acceptably self-tolerant repertoire. Nevertheless in this model the selected repertoire may still be reactive to self-peptides expressed heterogeneously in the thymus, and in particular to peptides expressed at high levels only on certain cell types. Van den Berg and Rand (100) review avidity-based models of ligand discrimination.

ALLOREACTIVITY AND MHC RESTRICTION

A high proportion (1–24%) of peripheral T cells are reactive to peptides presented on a foreign MHC allele (101–103), reflected clinically by acute T cell mediated rejection of grafts from MHC-mismatched donors. These promiscuous "allogenic" responses contrast with the low precursor frequency (10^{-6}–10^{-4}) in normal immune responses to peptides presented by an individual's own MHC. Allogenic responses are also apparently counter to the notion of MHC restriction. Reconciling these results may tell us great deal about the relative contributions of peptide and MHC binding motifs to the TCR signals driving selection, and how this breakdown influences the coverage and cross-reactivity of the T cell repertoire.

Detours and Perelson (82) used their digit-string model of TCR–pMHC interactions, described above, to show how the probabilities of responsiveness to self and foreign MHC emerge.

Mean alloreactivities of 1–2% arose naturally, at the lower end of the range of experimental estimates, and they showed that the alloreactivities of the pre- and post-selection repertoires are similar, as observed experimentally (17, 22). In essence, the modeling supports the hypothesis that the greater degree of alloreactivity than response frequency arises simply because many more pMHC ligands can be generated from one MHC than can be generated from one peptide (104). In other words, each TCR is triggered by ligands in a subset of pMHC shape space; one particular MHC along with its associated diversity of peptides will cover a far greater region of shape space than covered by one peptide and all the self-MHC alleles capable of presenting it; a given MHC will then stimulate far more of the T cell repertoire than will a given peptide.

They found that alloreactivity correlates with the extent of negative selection and inversely to the degree of MHC restriction. It can be seen intuitively how this emerges from their model. If negative selection is weak, positive selection must be correspondingly stringent in order to yield the selection efficiencies observed experimentally (3–5%). Stringent positive selection imposes an imprint of self-MHC on the repertoire – only those TCRs that bind strongly to self-MHC residues survive. The strength of binding to a randomly generated MHC not involved in selection (i.e., a foreign MHC) is then on average lower to that of self-MHC in the post-selection repertoire. This difference increases, and thus alloreactivity decreases, as the required strength of binding to self-MHC increases.

This trade-off between alloreactivity and restriction might be expected as they appear to be in conflict. However, experimental estimates of these two quantities are variable. The conclusions described above were derived analytically from a model capturing the mean-field behavior of the digit-string selection process, but did not deal with the variance in these measures of the repertoire outputs across specific simulations or experimental systems. The final study of the series (83) took a simulation-based approach, explicitly performing repertoire selection on random TCR and pMHC populations. This confirmed the inverse correlation between alloreactivity and MHC restriction and yielded sufficient variability to account for restriction ranging from absolute to partial in different settings.

Overall the digit-string model explored by Detours and colleagues yields remarkable agreement with many observations. Their model of TCR–pMHC binding is highly abstracted, but appears to be a powerful one. In part this might be because the relevant quantities for selection in their model are the minimum and maximum binding affinities that each TCR experiences during exposure to large samples of randomly generated pMHC strings. These two quantities will be drawn from extreme-value distributions, which should be insensitive to the distribution of binding strengths of randomly chosen TCR–pMHC pairs (70, 105). The additivity of the MHC and peptide contributions to the fate-determining signal is likely the most questionable assumption, as the authors point out. Fate decisions may be based on the sum of several TCR interactions (which means for example that positive selection may occur though proximal binding of multiple low-affinity ligands) and so an avidity-based model may be more appropriate. Another caveat is that the population-average model assumes that positive selection takes place on at most one MHC allele, which we will also return to.

INSIGHTS INTO FATE DETERMINATION MECHANISMS FROM STOCHASTICITY IN SELECTION

Regulatory T cells (Treg) are a distinct lineage of CD4SP cells thought to lie at the higher end of the spectrum of acceptable self-reactivity and play a crucial role in the control of autoimmunity and tolerance to innocuous antigens. Many experimental studies of Treg development have shown that cells with the same TCR can develop into conventional and regulatory T cells within the same selecting environment [see, for example, Ref. (58, 106)], illustrating again, as represented in so many models, the stochastic nature of selection. There are at least two possible sources of this stochasticity. In a purely selective model precursors with identical TCR might be predisposed to the conventional or Treg fates through natural variation in expression of factors involved in lineage commitment. In a purely instructive model, cells within a clone are uncommitted, and intra-clonal heterogeneity in fate may derive from variation in the experience of each thymocyte during selection – most likely because each encounters a different random sample of self-peptides.

Bains et al. (107) used a probabilistic, instructive model that reflects this view of fate determination driven entirely by antigenic experience during selection, in conjunction with data from Ref. (58) to infer the number of pMHC binding events involved in fate determination. In that study, the numbers of conventional and Treg cells with a transgenically expressed TCR were measured for varying abundances of that TCR's agonist peptide on thymic epithelial cells. Conventional cell numbers declined monotonically with agonist abundance, while Treg increased and then decreased. Thus as agonist abundance increased, it appeared that T cells were initially diverted into the Treg lineage, before the risk of deletion through exposure to agonist dominated. Using this information and a simple graphical argument they were able to infer that fate decisions could not be affinity-driven (that is, made on the basis of a single pMHC interaction) unless TCR sensitivity varies during development, for which there is evidence [see Ref. (107) and references therein]. This model also explains apparently paradoxical observations regarding the effect of partial and full TCR agonists on the efficiency of Treg production (108).

THE LIMITS OF NEGATIVE SELECTION

The potentially very large number of unique self-pMHC prompts the question of whether it is possible to tolerize thymocytes to all self-peptides within the timescale of thymic development. Müller and Bonhoeffer (109) studied this problem. Using constraints from the mouse proteome and the efficiencies of peptide production and binding to MHC, they estimated an upper limit of approximately 5×10^6 possible self-pMHC class I complexes. Notably, this diversity of self is several orders of magnitude lower than figures derived from the simple combinatoric arguments (64, 66) and is more closely aligned with an estimate that $\sim 10^5$ different nine-mers derived from the human proteome are expected to bind to one human MHC class I allele (73). The key quantity in Müller and Bonhoeffer's calculation is the probability P that a given self-pMHC is presented by any given APC in sufficient numbers for

negative selection to occur. The probability that a thymocyte specific for this (and only this) self-pMHC escapes negative selection is P_E in their notation – distinct from the probability of immune escape discussed above – and $P_E = (1 - P)^n$, where n is the number of unique APC encountered during selection. In this model, P_E is extremely sensitive to the number of copies h of a given self-pMHC that an APC needs to present in order to cause deletion – varying h between 15 and 1500 gives values of P_E between 10^{-11} and 0.8. Favoring the higher estimates of h, Müller and Bonhoeffer (109) concluded that negative selection on the potential diversity of self is likely to be very leaky. Instead, they suggest thymic selection operates on a restricted subset of self-pMHC, a constraint imposed by the number of APCs encountered during selection. This requires that further tolerogenic mechanisms operate in the periphery to prevent autoimmune response to self antigens not encountered in the thymus (53, 70).

To support their argument, Müller and Bonhoeffer (109) reverted to the older model of cross-reactivity and selection to generate another estimate of the number of selecting ligands using the observed efficiency of negative selection. Recall that the probability of thymocyte with cross-reactivity r escaping negative selection on N_s unique selecting ligands is $P = (1 - r)^{N_s} \simeq e^{-rN_s}$. Using the estimate of $r = 2 \times 10^{-5}$ (88), and $P \simeq 0.33$, they obtain $N_s \simeq 10^5$ unique selecting self-pMHC, or ~4% of the putative total number of self-pMHC. This estimate is consistent with those of Detours et al. (27). Both studies assume that this cross-reactivity r of thymocytes with self-pMHC is equal to the cross-reactivity of mature naive T cells to foreign pMHC. Since negative selection likely acts as a filter to reduce cross-reactivity in the post-selection repertoire (see above), this assumption is moot. But the need to meet the empirical constraint $e^{-rN_s} \simeq 0.33$ implies that higher values of r would reduce the number of unique selecting ligands N_s even further.

A subsequent exchange (110, 111) discussed the assumption that each TCR negatively selects only on a single self-pMHC ligand. Müller and Bonhoeffer (111) argued that in the Bernoulli trial model of cross-reactivity and selection, a 33% probability of survival implies that another third of all thymocytes were reactive to one self-pMHC only, giving some quantitative support to their original model. The discussion also addressed whether N_s is constrained by the residence time in the thymus or is a result of restricted presentation of self antigens. Müller and Bonhoeffer (111) favored the latter, presuming that evolution has optimized the thymic residence time for the purposes of efficient selection on a subset of self-peptides. More recently it has been argued that incomplete depletion of self-reactive cells in the thymus may be sufficient for robust self/non-self discrimination in the periphery, if interactions facilitating consensus between T cells are required for the initiation or suppression of immune responses (70).

OPTIMALITY OF INDIVIDUAL MHC DIVERSITY – CONSTRAINTS ARISING FROM THYMIC SELECTION

The polymorphism of the MHC is huge, with hundreds of alleles identified at the HLA-A, HLA-B, and HLA-DR loci in humans (MHC is referred to as HLA in humans but hereon the term MHC is generally used, for simplicity). This diversification is thought not to have occurred by genetic drift but by two non-exclusive mechanisms. Heterozygote advantage (112, 113) suggests that individuals expressing more unique MHC alleles gain fitness by being able to present a larger array of pathogen peptides. Overall the evidence for heterozygote advantage in experimental models of infection is equivocal, though, and it has been argued with a quantitative model that this mechanism alone is insufficient to explain the extent of allelic diversity (114). Another theory is that MHC polymorphism is maintained by frequency-dependent selection under pathogen pressure, in which rare alleles confer protection against pathogen subversion of peptide presentation by commonly expressed alleles (115).

Intriguingly, individuals possess only a small proportion of all MHC alleles. Heterozygous humans possess six at the major HLA-A, HLA-B, and HLA-C loci, which code for MHC class I molecules that present peptides to $CD8^+$ T cells, and six to eight at the HLA-DP, HLA-DQ, and HLA-DR MHC class II loci, which present to $CD4^+$ T cells. A common explanation for this restricted within-individual diversity is that it derives from the need to generate a broad, functional, and self-tolerant TCR repertoire in the thymus without excessive negative selection (116, 117). The qualitative argument is as follows. If n is the number of MHC alleles per person, then increasing n both increases the diversity of pathogen-derived peptides that can be presented and increases the probability that a thymocyte will be able to obtain positively-selecting signals. On the other hand, higher n will also increase the range of self-peptides that can be presented. This will increase the stringency of negative selection, leading to inefficient generation of T cells in the thymus and potential gaps in the repertoire's coverage of peptide space. The observed number of different MHC molecules per individual may result from a trade-off between these demands.

The nature of MHC restriction needs to be considered carefully in these arguments. If restriction is absolute and each TCR recognizes only one MHC allele, increasing the number of alleles per person simply increases the size and diversity of the T cell repertoire with no cost because selection operates on each MHC-restricted subset of the pre-selection repertoire independently. In this case an upper limit to within-host MHC diversity might derive only from the need for APC to display sufficient numbers of peptides in conjunction with each MHC molecule to reliably mediate selection or immune activation. The trade-off evident in the qualitative argument above arises when MHC restriction is not absolute and thymocytes are capable of being positively and/or negatively selected on more than one allele.

Woelfing et al. (118) provide an excellent review of theoretical approaches to understanding intra-individual MHC diversity, but we outline the key results here. Nowak et al. (119) were the first to assess the qualitative trade-off argument using a mathematical model. In their analysis they defined h and f to be the proportions of T cells capable of being positively and negatively selected, respectively, by a given MHC allele. If an individual expresses n distinct MHC alleles, they argue that the proportion of the T cell repertoire surviving selection is

$$(1 - (1 - h)^n)(1 - f)^n.$$

The first term represents positive selection; $(1 - h)^n$ is the probability that a TCR fails to be selected by any MHC. The second term

represents negative selection; $(1-f)^n$ is the probability that a TCR is not negatively selected by any MHC. The proportion of the repertoire surviving is maximized at $n=(1/h)\log(1+h/f)$. They argue that $h \leq f$, supported by the experimental and modeling consensus is that positive selection is more stringent than negative selection. This gives $n \sim 1/f$. However, using only the assumptions that $hn \ll 1$, or that it is rare for a TCR to be positively selected on more than one MHC allele, and that the proportion of all peptides that can bind to a given MHC is $\ll 1$, they calculate that $n=2/f$ maximizes the probability of a response to a randomly chosen foreign pMHC.

Borghans et al. (120) pointed out that this model contains an inconsistency, which allows for cells that fail to be positively selected on one MHC to be negatively selected by the same MHC. They denoted p and n to be the unconditional probabilities that one TCR is positively and negatively selected by a given MHC molecule. Then $n < p$, because the number of cells that fail negative selection on one MHC is necessarily smaller than the number that audition for it following positive selection on that same MHC. The proportion of the original repertoire that survives is then

$$(1-n)^M - (1-p)^M. \tag{4}$$

This model effectively lowers the stringency of negative selection expressed in Nowak et al. (119) and so reduces the cost of increasing the number of MHC alleles. They estimated the probabilities p and n were 0.01 and 0.005 respectively, using the known efficiencies of positive and negative selection in mice with known numbers of MHC alleles. The optimal value of M for these parameter values is far larger than observed allele numbers; conversely, asking what values of p and n correspond to the observed ranges of M being optimal leads to unrealistic levels of positive and negative selection. Their analysis therefore questions the trade-off hypothesis as an explanation of limited MHC diversity.

They suggest alternatives. They estimate that existing typical numbers of MHC alleles together with TCR cross-reactivity may be "good enough" for maximizing the probability of responding to a foreign peptide on self-MHC – in this case the selective pressure for increasing MHC alleles is weak or absent. Alternatively, increased numbers of MHC alleles may increase the risk of autoimmunity through cross-reactivity of T cells responding to antigen that have not been fully tolerized to self. Finally, limited numbers of MHC alleles may allow for sufficient densities of pMHC on the surface of antigen-presenting cells to be able to efficiently select and activate MHC-restricted T cells.

MHC restriction is not absolute in the models described above, although it holds approximately for positive selection when the per-allele positive selection probability p is small. However, there is evidence to suggest that MHC restriction is not manifest strongly at the positive selection stage. Zerrahn et al. (22) observed that a relatively large proportion of TCR still positively select when a single type of pMHC was expressed in the thymus. In that study, pre-selection TCRs had approximately a 5% chance of responding to a given class II MHC, independently for different alleles, validating one of the assumptions of these simple probabilistic selection models. On similar lines, Huseby et al. (92) found that the positively-selected repertoire contains TCR with a high degree of cross-reactivity across MHC alleles, and suggested that MHC restriction emerges as a result of negative selection. Finally, the high degree of alloreactivity suggests that positive selection is only weakly MHC-restricted, and that failure to positive select reflects a generic inability to bind to MHC.

Motivated by this possibility, Woelfing et al. (118) revisited these probabilistic models. They assumed positive selection is highly degenerate with respect to MHC and that even very weak cross-reactivity with any allele is sufficient. Under this assumption, one of the presumed advantages of high MHC diversity is removed. Maximizing the probability of mounting an immune response, they estimated the optimal MHC diversity to be in a physiological range of 3–25.

Van den Berg and Rand (121) used a very different and sophisticated approach to the same optimality problem using a mechanistic, stochastic model of TCR triggering rather than the probabilistic repertoire-based models described above. Considering negative selection only, they concluded that limited individual MHC diversity is beneficial for self–non-self discrimination. The essence of their mathematical argument is that restricting the "diversity of foreign" is the key to increasing the signal-to-noise ratio for aTCR attempting to discriminate a foreign peptide from the background of self. This is achieved with a combination of limiting the number of MHC alleles each TCR can recognize (MHC restriction) and limiting the number of peptides that can be presented from one protein on one MHC allele ("peptide selectivity") to be roughly one. However, the need to ensure that every foreign protein is represented requires multiple MHC alleles, placing a theoretical lower bound on their number. An upper bound comes from the requirement that the density of relevant pMHC ligands must not fall too low on the surface of an APC, similar to the suggestion in Borghans et al. (120) – if a given pMHC is diluted by too many MHC, the relevant TCR will experience fluctuations in signaling that may reduce its discriminatory power. They conclude that of the order 10 MHC alleles is optimal. Notably, as in Ref. (118), this estimate arises without any constraints from positive selection.

SUMMARY

This review has outlined how several relatively simple descriptions of single TCR–pMHC interactions have been used to understand aspects of TCR repertoire development. However, the discussion is necessarily incomplete. In particular, there is an extensive literature exploring the molecular mechanisms by which individual or collections of TCR discriminate between ligands of different affinities [see, for example, Ref. (100, 122–126)], which has direct relevance to thymic selection. It remains unclear how proximal TCR signals derived from multiple and diverse pMHC ligands can drive the emergence of specificity and MHC restriction in the post-selection repertoire, although the models of selection on ensembles of ligands have made steps in this direction (99, 121). Are repeated super-threshold contacts required for negative selection, or is a single encounter with a high affinity ligand sufficient to cause deletion?

Many of the models discussed here assume that a single interaction above a minimum signaling threshold is sufficient for positive selection. However, there is evidence that repeated or sustained TCR signaling is required during the DP stage for positive selection

to occur [see, for example, Ref. (17, 127)]. This may explain findings that positive and negative selection take place concurrently (46).

Overall it is remarkable how much insight into the quantitative aspects of thymic selection has emerged from highly abstracted models. However, there remain a lot of open areas for research, and many of the questions raised in the introduction are still unresolved. Regulatory T cell development in particular has received very little attention from modelers, and already it appears that the simplest extension to the simple probabilistic fixed-threshold model to include a fixed range of affinity or avidity for Treg selection is not sufficient to explain many experimental observations (107). The task of synthesizing and reconciling the huge diversity of experimental data related to thymic development is a daunting one, but the information available is perhaps currently underexploited by theorists.

ACKNOWLEDGMENTS

The author thanks Charles Sinclair and the reviewers for helpful comments. This work was supported by the NIH (R01AI093870).

REFERENCES

1. Paul WE, editor. *Fundamental Immunology*. 6 ed. Philadelphia: Lippincott Williams & Wilkins (2008).
2. Starr TK, Jameson SC, Hogquist KA. Positive and negative selection of T cells. *Annu Rev Immunol* (2003) **21**:139–76. doi:10.1146/annurev.immunol.21.120601.141107
3. Moran AE, Hogquist KA. T-cell receptor affinity in thymic development. *Immunology* (2012) **135**(4):261–7. doi:10.1111/j.1365-2567.2011.03547.x
4. Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, Shortman K. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *J Exp Med* (1993) **178**(2):615–22. doi:10.1084/jem.178.2.615
5. Scollay RG, Butcher EC, Weissman IL. Thymus cell migration. Quantitative aspects of cellular traffic from the thymus to the periphery in mice. *Eur J Immunol* (1980) **10**(3):210–8. doi:10.1002/eji.1830100310
6. Egerton M, Scollay R, Shortman K. Kinetics of mature T-cell development in the thymus. *Proc Natl Acad Sci U S A* (1990) **87**(7):2579–82. doi:10.1073/pnas.87.7.2579
7. Graziano M, St-Pierre Y, Beauchemin C, Desrosiers M, Potworowski EF. The fate of thymocytes labeled in vivo with CFSE. *Exp Cell Res* (1998) **240**(1):75–85. doi:10.1006/excr.1997.3900
8. Itano A, Robey E. Highly efficient selection of CD4 and CD8 lineage thymocytes supports an instructive model of lineage commitment. *Immunity* (2000) **12**(4):383–9. doi:10.1016/S1074-7613(00)80190-9
9. Quackenbush RC, Shields AF. Local re-utilization of thymidine in normal mouse tissues as measured with iododeoxyuridine. *Cell Tissue Kinet* (1988) **21**(6):381–7.
10. McCaughtry TM, Wilken MS, Hogquist KA. Thymic emigration revisited. *J Exp Med* (2007) **204**(11):2513–20. doi:10.1084/jem.20070601
11. Sinclair C, Bains I, Yates AJ, Seddon B. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proc Natl Acad Sci U S A* (2013) **110**(31):E2905–14. doi:10.1073/pnas.1304859110
12. Stritesky GL, Xing Y, Erickson JR, Kalekar LA, Wang X, Mueller DL, et al. Murine thymic selection quantified using a unique method to capture deleted T cells. *Proc Natl Acad Sci U S A* (2013) **110**(12):4679–84. doi:10.1073/pnas.1217532110
13. Thomas-Vaslin V, Altes HK, de Boer RJ, Klatzmann D. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *J Immunol* (2008) **180**(4):2240–50.
14. Shortman K, Vremec D, Egerton M. The kinetics of T cell antigen receptor expression by subgroups of CD4+8+ thymocytes: delineation of CD4+8+3(2+) thymocytes as post-selection intermediates leading to mature T cells. *J Exp Med* (1991) **173**(2):323–32. doi:10.1084/jem.173.2.323
15. Huesmann M, Scott B, Kisielow P, von Boehmer H. Kinetics and efficacy of positive selection in the thymus of normal and T cell receptor transgenic mice. *Cell* (1991) **66**(3):533–40. doi:10.1016/0092-8674(81)90016-7
16. von Boehmer H. Positive selection of lymphocytes. *Cell* (1994) **76**(2):219–28. doi:10.1016/0092-8674(94)90330-1
17. Merkenschlager M, Graf D, Lovatt M, Bommhardt U, Zamoyska R, Fisher AG. How many thymocytes audition for selection? *J Exp Med* (1997) **186**(7):1149–58. doi:10.1084/jem.186.7.1149
18. Monteiro MC, Couceiro S, Penha-Gonçalves C. The multigenic structure of the MHC locus contributes to positive selection efficiency: a role for MHC class II gene-specific restriction. *Eur J Immunol* (2005) **35**(12):3622–30. doi:10.1002/eji.200535190
19. Ignatowicz L, Rees W, Pacholczyk R, Ignatowicz H, Kushnir E, Kappler J, et al. T cells can be activated by peptides that are unrelated in sequence to their selecting peptide. *Immunity* (1997) **7**(2):179–86. doi:10.1016/S1074-7613(00)80521-X
20. van Meerwijk JP, Marguerat S, Lees RK, Germain RN, Fowlkes BJ, MacDonald HR. Quantitative impact of thymic clonal deletion on the T cell repertoire. *J Exp Med* (1997) **185**(3):377–83. doi:10.1084/jem.185.3.377
21. Tourne S, Miyazaki T, Oxenius A, Klein L, Fehr T, Kyewski B, et al. Selection of a broad repertoire of CD4+ T cells in H-2Ma0/0 mice. *Immunity* (1997) **7**(2):187–95. doi:10.1016/S1074-7613(00)80522-1
22. Zerrahn J, Held W, Raulet DH. The MHC reactivity of the T cell repertoire prior to positive and negative selection. *Cell* (1997) **88**(5):627–36. doi:10.1016/S0092-8674(00)81905-4
23. Hale JS, Boursalian TE, Turk GL, Fink PJ. Thymic output in aged mice. *Proc Natl Acad Sci U S A* (2006) **103**(22):8447–52. doi:10.1073/pnas.0601040103
24. Steinmann GG, Klaus B, Muller-Hermelink HK. The involution of the ageing human thymic epithelium is independent of puberty. A morphometric study. *Scand J Immunol* (1985) **22**(5):563–75. doi:10.1111/j.1365-3083.1985.tb01916.x
25. Scollay R, Godfrey DI. Thymic emigration: conveyor belts or lucky dips? *Immunol Today* (1995) **16**(6):268–73. doi:10.1016/0167-5699(95)80179-0
26. Porritt HE, Gordon K, Petrie HT. Kinetics of steady-state differentiation and mapping of intrathymic-signaling environments by stem cell transplantation in nonirradiated mice. *J Exp Med* (2003) **198**(6):957–62. doi:10.1084/jem.20030837
27. Detours V, Mehr R, Perelson AS. Deriving quantitative constraints on T cell selection from data on the mature T cell repertoire. *J Immunol* (2000) **164**(1):121–8.
28. Lucas B, Vasseur F, Penit C. Production, selection, and maturation of thymocytes with high surface density of TCR. *J Immunol* (1994) **153**(1):53–62.
29. Hare KJ, Wilkinson RW, Jenkinson EJ, Anderson G. Identification of a developmentally regulated phase of postselection expansion driven by thymic epithelium. *J Immunol* (1998) **160**(8):3666–72.
30. Pénit C, Vasseur F. Expansion of mature thymocyte subsets before emigration to the periphery. *J Immunol* (1997) **159**(10):4848–56.
31. Okamoto Y, Douek DC, McFarland RD, Koup RA. Effects of exogenous interleukin-7 on human thymus function. *Blood* (2002) **99**(8):2851–8. doi:10.1182/blood.V99.8.2851
32. Junge S, Kloeckener-Gruissem B, Zufferey R, Keisker A, Salgo B, Fauchere J-C, et al. Correlation between recent thymic emigrants and CD31+ (PECAM-1) CD4+ T cells in normal individuals during aging and in lymphopenic children. *Eur J Immunol* (2007) **37**(11):3270–80. doi:10.1002/eji.200636976
33. Saini M, Sinclair C, Marshall D, Tolaini M, Sakaguchi S, Seddon B. Regulation of Zap70 expression during thymocyte development enables temporal separation of CD4 and CD8 repertoire selection at different signaling thresholds. *Sci Signal* (2010) **3**(114):ra23. doi:10.1126/scisignal.2000702
34. Liu X, Adams A, Wildt KF, Aronow B, Feigenbaum L, Bosselut R. Restricting Zap70 expression to CD4+CD8+ thymocytes reveals a T cell receptor-dependent proofreading mechanism controlling the completion of positive selection. *J Exp Med* (2003) **197**(3):363–73. doi:10.1084/jem.20021698
35. Dzhagalov IL, Chen KG, Herzmark P, Robey EA. Elimination of self-reactive T cells in the thymus: a timeline for negative selection. *PLoS Biol* (2013) **11**(5):e1001566. doi:10.1371/journal.pbio.1001566
36. Mehr R, Globerson A, Perelson AS. Modeling positive and negative selection and differentiation processes in the thymus. *J Theor Biol* (1995) **175**(1):103–26. doi:10.1006/jtbi.1995.0124

37. Mehr R, Perelson AS, Fridkis-Hareli M, Globerson A. Feedback regulation of T cell development in the thymus. *J Theor Biol* (1996) **181**(2):157–67. doi:10.1006/jtbi.1996.0122
38. Mehr R, Abel L, Ubezio P, Globerson A, Agur Z. A mathematical model of the effect of aging on bone marrow cells colonizing the thymus. *Mech Ageing Dev* (1993) **67**(1–2):159–72. doi:10.1016/0047-6374(93)90120-G
39. Mehr R, Segel L, Sharp A, Globerson A. Colonization of the thymus by T cell progenitors: models for cell-cell interactions. *J Theor Biol* (1994) **170**(3):247–57. doi:10.1006/jtbi.1994.1185
40. Mehr R, Fridkis-Hareli M, Abel L, Segel L, Globerson A. Lymphocyte development in irradiated thymuses: dynamics of colonization by progenitor cells and regeneration of resident cells. *J Theor Biol* (1995) **177**(2):181–92. doi:10.1006/jtbi.1995.0237
41. Witt CM, Raychaudhuri S, Schaefer B, Chakraborty AK, Robey EA. Directed migration of positively selected thymocytes visualized in real time. *PLoS Biol* (2005) **3**(6):e373. doi:10.1371/journal.pbio.0030373
42. Surh CD, Sprent J. T-cell apoptosis detected in situ during positive and negative selection in the thymus. *Nature* (1994) **372**(6501):100–3. doi:10.1038/372100a0
43. Sprent J, Kishimoto H. The thymus and negative selection. *Immunol Rev* (2002) **185**:126–35. doi:10.1034/j.1600-065X.2002.18512.x
44. Le Borgne M, Ladi E, Dzhagalov I, Herzmark P, Liao YF, Chakraborty AK, et al. The impact of negative selection on thymocyte migration in the medulla. *Nat Immunol* (2009) **10**(8):823–30. doi:10.1038/ni.1761
45. McCaughtry TM, Baldwin TA, Wilken MS, Hogquist KA. Clonal deletion of thymocytes can occur in the cortex with no involvement of the medulla. *J Exp Med* (2008) **205**(11):2575–84. doi:10.1084/jem.20080866
46. Baldwin KK, Trenchak BP, Altman JD, Davis MM. Negative selection of T cells occurs throughout thymic development. *J Immunol* (1999) **163**(2):689–98.
47. Lorenz RG, Allen PM. Thymic cortical epithelial cells lack full capacity for antigen presentation. *Nature* (1989) **340**(6234):557–9. doi:10.1038/340557a0
48. Mizuochi T, Kasai M, Kokuho T, Kakiuchi T, Hirokawa K. Medullary but not cortical thymic epithelial cells present soluble antigens to helper T cells. *J Exp Med* (1992) **175**(6):1601–5. doi:10.1084/jem.175.6.1601
49. Gray DHD, Seach N, Ueno T, Milton MK, Liston A, Lew AM, et al. Developmental kinetics, turnover, and stimulatory capacity of thymic epithelial cells. *Blood* (2006) **108**(12):3777–85. doi:10.1182/blood-2006-02-004531
50. Hogquist KA, Xing Y. Why CD8+ T cells need diversity when growing up. *Immunity* (2010) **32**(1):5–6. doi:10.1016/j.immuni.2010.01.005
51. Nitta T, Murata S, Sasaki K, Fujii H, Ripen AM, Ishimaru N, et al. Thymoproteasome shapes immunocompetent repertoire of CD8+ T cells. *Immunity* (2010) **32**(1):29–40. doi:10.1016/j.immuni.2009.10.009
52. Faro J, Velasco S, González-Fernández A, Bandeira A. The impact of thymic antigen diversity on the size of the selected T cell repertoire. *J Immunol* (2004) **172**(4):2247–55.
53. Mason D. Some quantitative aspects of T-cell repertoire selection: the requirement for regulatory T cells. *Immunol Rev* (2001) **182**:80–8. doi:10.1034/j.1600-065X.2001.1820106.x
54. Borgulya P, Kishi H, Müller U, Kirberg J, von Boehmer H. Development of the CD4 and CD8 lineage of T cells: instruction versus selection. *EMBO J* (1991) **10**(4):913–8.
55. Corbella P, Moskophidis D, Spanopoulou E, Mamalaki C, Tolaini M, Itano A, et al. Functional commitment to helper T cell lineage precedes positive selection and is independent of T cell receptor MHC specificity. *Immunity* (1994) **1**(4):269–76. doi:10.1016/1074-7613(94)90078-7
56. Itano A, Kioussis D, Robey E. Stochastic component to development of class I major histocompatibility complex-specific T cells. *Proc Natl Acad Sci U S A* (1994) **91**(1):220–4. doi:10.1073/pnas.91.1.220
57. Germain RN. T-cell development and the CD4-CD8 lineage decision. *Nat Rev Immunol* (2002) **2**(5):309–22. doi:10.1038/nri798
58. van Santen H-M, Benoist C, Mathis D. Number of T reg cells that differentiate does not increase upon encounter of agonist ligand on thymic epithelial cells. *J Exp Med* (2004) **200**(10):1221–30. doi:10.1084/jem.20041022
59. Bautista JL, Lio C-WJ, Lathrop SK, Forbush K, Liang Y, Luo J, et al. Intraclonal competition limits the fate determination of regulatory T cells in the thymus. *Nat Immunol* (2009) **10**(6):610–7. doi:10.1038/ni.1739
60. Lee H-M, Bautista JL, Scott-Browne J, Mohan JF, Hsieh C-S. A broad range of self-reactivity drives thymic regulatory T cell selection to limit responses to self. *Immunity* (2012) **37**(3):475–86. doi:10.1016/j.immuni.2012.07.009
61. Souza-e Silva H, Savino W, Feijóo RA, Vasconcelos ATR. A cellular automata-based mathematical model for thymocyte development. *PLoS One* (2009) **4**(12):e8233. doi:10.1371/journal.pone.0008233
62. Liu X, Bosselut R. Duration of TCR signaling controls CD4-CD8 lineage differentiation in vivo. *Nat Immunol* (2004) **5**(3):280–8. doi:10.1038/ni1040
63. Efroni S, Harel D, Cohen IR. Emergent dynamics of thymocyte development and lineage determination. *PLoS Comput Biol* (2007) **3**(1):e13. doi:10.1371/journal.pcbi.0030013
64. Mason D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol Today* (1998) **19**(9):395–404. doi:10.1016/S0167-5699(98)01299-7
65. Ishizuka J, Grebe K, Shenderov E, Peters B, Chen Q, Peng Y, et al. Quantitating T cell cross-reactivity for unrelated peptide antigens. *J Immunol* (2009) **183**(7):4337–45. doi:10.4049/jimmunol.0901607
66. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol* (2012) **12**(9):669–77. doi:10.1038/nri3279
67. De Boer RJ, Perelson AS. How diverse should the immune system be? *Proc Biol Sci* (1993) **252**(1335):171–5. doi:10.1098/rspb.1993.0062
68. Whitaker L, Renton AM. On the plausibility of the clonal expansion theory of the immune system in terms of the combinatorial possibilities of amino-acids in antigen and self-tolerance. *J Theor Biol* (1993) **164**(4):531–6. doi:10.1006/jtbi.1993.1171
69. Nemazee D. Antigen receptor 'capacity' and the sensitivity of self-tolerance. *Immunol Today* (1996) **17**(1):25–9. doi:10.1016/0167-5699(96)80565-2
70. Butler TC, Kardar M, Chakraborty AK. Quorum sensing allows T cells to discriminate between self and nonself. *Proc Natl Acad Sci U S A* (2013) **110**(29):11833–8. doi:10.1073/pnas.1222467110
71. Borghans JA, De Boer RJ. Crossreactivity of the T-cell receptor. *Immunol Today* (1998) **19**(9):428–9. doi:10.1016/S0167-5699(98)01317-6
72. Percus JK, Percus OE, Perelson AS. Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proc Natl Acad Sci U S A* (1993) **90**(5):1691–5. doi:10.1073/pnas.90.5.1691
73. Burroughs NJ, de Boer RJ, Kesmir C. Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics* (2004) **56**(5):311–20. doi:10.1007/s00251-004-0691-0
74. Grossman Z, Singer A. Tuning of activation thresholds explains flexibility in the selection and development of T cells in the thymus. *Proc Natl Acad Sci U S A* (1996) **93**(25):14747–52. doi:10.1073/pnas.93.25.14747
75. Anderton SM, Wraith DC. Selection and fine-tuning of the autoimmune T-cell repertoire. *Nat Rev Immunol* (2002) **2**(7):487–98. doi:10.1038/nri842
76. Scherer A, Noest A, de Boer RJ. Activation-threshold tuning in an affinity model for the T-cell repertoire. *Proc Biol Sci* (2004) **271**(1539):609–16. doi:10.1098/rspb.2003.2653
77. Cox AL, Skipper J, Chen Y, Henderson RA, Darrow TL, Shabanowitz J, et al. Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines. *Science* (1994) **264**(5159):716–9. doi:10.1126/science.7513441
78. Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K, et al. Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* (1992) **256**(5065):1817–20. doi:10.1126/science.1319610
79. Borghans JA, Noest AJ, De Boer RJ. How specific should immunological memory be? *J Immunol* (1999) **163**(2):569–75.
80. Borghans JAM, De Boer RJ. Memorizing innate instructions requires a sufficiently specific adaptive immune system. *Int Immunol* (2002) **14**(5):525–32. doi:10.1093/intimm/14.5.525
81. Detours V, Sulzer B, Perelson AS. Size and connectivity of the idiotypic network are independent of the discreteness of the affinity distribution. *J Theor Biol* (1996) **183**(4):409–16. doi:10.1006/jtbi.1996.0231
82. Detours V, Perelson AS. Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proc Natl Acad Sci U S A* (1999) **96**(9):5153–8. doi:10.1073/pnas.96.9.5153
83. Detours V, Perelson AS. The paradox of alloreactivity and self MHC restriction: quantitative analysis and statistics. *Proc Natl Acad Sci U S A* (2000) **97**(15):8479–83. doi:10.1073/pnas.97.15.8479
84. Perelson AS, Weisbuch G. Immunology for physicists. *Rev Mod Phys* (1997) **69**:1219–68. doi:10.1103/RevModPhys.69.1219
85. Stockinger H, Pfizenmaier K, Hardt C, Rodt H, Röllinghoff M, Wagner H. H-2 restriction as a consequence of intentional priming: T cells of fully allogeneic

chimeric mice as well as of normal mice respond to foreign antigens in the context of H-2 determinants not encountered on thymic epithelial cells. *Proc Natl Acad Sci U S A* (1980) **77**(12):7390–4. doi:10.1073/pnas.77.12.7390
86. Merkenschlager M, Terry L, Edwards R, Beverley PC. Limiting dilution analysis of proliferative responses in human lymphocyte populations defined by the monoclonal antibody UCHL1: implications for differential CD45 expression in T cell memory formation. *Eur J Immunol* (1988) **18**(11):1653–61. doi:10.1002/eji.1830181102
87. Zinkernagel RM. Immunology taught by viruses. *Science* (1996) **271**(5246): 173–8. doi:10.1126/science.271.5246.173
88. Blattman JN, Antia R, Sourdive DJD, Wang X, Kaech SM, Murali-Krishna K, et al. Estimating the precursor frequency of naive antigen-specific CD8 T cells. *J Exp Med* (2002) **195**(5):657–64. doi:10.1084/jem.20001021
89. Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, et al. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* (2007) **27**(2):203–13. doi:10.1016/j.immuni.2007.07.007
90. Huseby ES, Crawford F, White J, Kappler J, Marrack P. Negative selection imparts peptide specificity to the mature T cell repertoire. *Proc Natl Acad Sci U S A* (2003) **100**(20):11565–70. doi:10.1073/pnas.1934636100
91. Slifka MK, Blattman JN, Sourdive DJD, Liu F, Huffman DL, Wolfe T, et al. Preferential escape of subdominant CD8+ T cells during negative selection results in an altered antiviral T cell hierarchy. *J Immunol* (2003) **170**(3):1231–9.
92. Huseby ES, White J, Crawford F, Vass T, Becker D, Pinilla C, et al. How the T cell repertoire becomes peptide and MHC specific. *Cell* (2005) **122**(2):247–60. doi:10.1016/j.cell.2005.05.013
93. Chao DL, Davenport MP, Forrest S, Perelson AS. The effects of thymic selection on the range of T cell cross-reactivity. *Eur J Immunol* (2005) **35**(12):3452–9. doi:10.1002/eji.200535098
94. Kraj P, Pacholczyk R, Ignatowicz L. Alpha beta TCRs differ in the degree of their specificity for the positively selecting MHC/peptide ligand. *J Immunol* (2001) **166**(4):2251–9.
95. Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci U S A* (2008) **105**(43):16671–6. doi:10.1073/pnas.0808081105
96. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* (1996) **256**(3):623–44. doi:10.1006/jmbi.1996.0114
97. Košmrlj A, Chakraborty AK, Kardar M, Shakhnovich EI. Thymic selection of T-cell receptors as an extreme value problem. *Phys Rev Lett* (2009) **103**:068103. doi:10.1103/PhysRevLett.103.068103
98. Kosmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, et al. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* (2010) **465**(7296):350–4. doi:10.1038/nature08997
99. van den Berg HA, Rand DA, Burroughs NJ. A reliable and safe T cell repertoire based on low-affinity T cell receptors. *J Theor Biol* (2001) **209**(4):465–86. doi:10.1006/jtbi.2001.2281
100. van den Berg HA, Rand DA. Quantitative theories of T-cell responsiveness. *Immunol Rev* (2007) **216**:81–92. doi:10.1111/j.1600-065X.2006.00491.x
101. Bevan MJ, Langman RE, Cohn M. H-2 antigen-specific cytotoxic T cells induced by concanavalin A: estimation of their relative frequency. *Eur J Immunol* (1976) **6**(3):150–6. doi:10.1002/eji.1830060303
102. Ashwell JD, Chen C, Schwartz RH. High frequency and nonrandom distribution of alloreactivity in T cell clones selected for recognition of foreign antigen in association with self class II molecules. *J Immunol* (1986) **136**(2): 389–95.
103. Suchin EJ, Langmuir PB, Palmer E, Sayegh MH, Wells AD, Turka LA. Quantifying the frequency of alloreactive T cells in vivo: new answers to an old question. *J Immunol* (2001) **166**(2):973–81.
104. Matzinger P, Bevan MJ. Hypothesis: why do so many lymphocytes respond to major histocompatibility antigens? *Cell Immunol* (1977) **29**(1):1–5. doi:10.1016/0008-8749(77)90269-6
105. Fisher RA, Tippet LHC. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc Camb Philos Soc* (1928) **24**:180–90. doi:10.1017/S0305004100015681
106. Picca CC, Oh S, Panarey L, Aitken M, Basehoar A, Caton AJ. Thymocyte deletion can bias Treg formation toward low-abundance self-peptide. *Eur J Immunol* (2009) **39**(12):3301–6. doi:10.1002/eji.200939709
107. Bains I, van Santen HM, Seddon B, Yates AJ. Models of self-peptide sampling by developing T cells identify candidate mechanisms of thymic selection. *PLoS Comput Biol* (2013) **9**(7):e1003102. doi:10.1371/journal.pcbi.1003102
108. Picca CC, Simons DM, Oh S, Aitken M, Perng OA, Mergenthaler C, et al. CD4+ CD25+ Foxp3+ regulatory T cell formation requires more specific recognition of a self-peptide than thymocyte deletion. *Proc Natl Acad Sci U S A* (2011) **108**(36):14890–5. doi:10.1073/pnas.1103810108
109. Müller V, Bonhoeffer S. Quantitative constraints on the scope of negative selection. *Trends Immunol* (2003) **24**(3):132–5. doi:10.1016/S1471-4906(03)00028-0
110. Bandeira A, Faro J. Quantitative constraints on the scope of negative selection: robustness and weaknesses. *Trends Immunol* (2003) **24**(4):172–3. doi:10.1016/S1471-4906(03)00055-3
111. Müller V, Bonhoeffer S. Response to Bandeira and Faro: closing the circle of constraints. *Trends Immunol* (2003) **24**(4):173–5. doi:10.1016/S1471-4906(03)00056-5
112. Doherty PC, Zinkernagel RM. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* (1975) **256**(5512):50–2. doi:10.1038/256050a0
113. McClelland EE, Penn DJ, Potts WK. Major histocompatibility complex heterozygote superiority during coinfection. *Infect Immun* (2003) **71**(4):2079–86. doi:10.1128/IAI.71.4.2079-2086.2003
114. De Boer RJ, Borghans JAM, van Boven M, Kesmir C, Weissing FJ. Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. *Immunogenetics* (2004) **55**(11):725–31. doi:10.1007/s00251-003-0629-y
115. Borghans JAM, Beltman JB, De Boer RJ. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* (2004) **55**(11):732–9. doi:10.1007/s00251-003-0630-5
116. Vidovic D, Matzinger P. Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* (1988) **336**(6196):222–5. doi:10.1038/336222a0
117. Lawlor DA, Zemmour J, Ennis PD, Parham P. Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. *Annu Rev Immunol* (1990) **8**:23–63. doi:10.1146/annurev.iy.08.040190.000323
118. Woelfing B, Traulsen A, Milinski M, Boehm T. Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philos Trans R Soc Lond B Biol Sci* (2009) **364**(1513):117–28. doi:10.1098/rstb.2008.0174
119. Nowak MA, Tarczy-Hornoch K, Austyn JM. The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci U S A* (1992) **89**(22):10896–9. doi:10.1073/pnas.89.22.10896
120. Borghans JAM, Noest AJ, De Boer RJ. Thymic selection does not limit the individual MHC diversity. *Eur J Immunol* (2003) **33**(12):3353–8. doi:10.1002/eji.200324365
121. van den Berg HA, Rand DA. Antigen presentation on MHC molecules as a diversity filter that enhances immune efficacy. *J Theor Biol* (2003) **224**(2):249–67. doi:10.1016/S0022-5193(03)00162-0
122. Altan-Bonnet G, Germain RN. Modeling T cell antigen discrimination based on feedback control of digital ERK responses. *PLoS Biol* (2005) **3**(11):e356. doi:10.1371/journal.pbio.0030356
123. Prasad A, Zikherman J, Das J, Roose JP, Weiss A, Chakraborty AK. Origin of the sharp boundary that discriminates positive and negative selection of thymocytes. *Proc Natl Acad Sci U S A* (2009) **106**(2):528–33. doi:10.1073/pnas.0805981105
124. Palmer E, Naeher D. Affinity threshold for thymic selection through a T-cell receptor-co-receptor zipper. *Nat Rev Immunol* (2009) **9**(3):207–13. doi:10.1038/nri2469
125. Govern CC, Paczosa MK, Chakraborty AK, Huseby ES. Fast on-rates allow short dwell time ligands to activate T cells. *Proc Natl Acad Sci U S A* (2010) **107**(19):8724–9. doi:10.1073/pnas.1000966107
126. Currie J, Castro M, Lythe G, Palmer E, Molina-París C. A stochastic T cell response criterion. *J R Soc Interface* (2012) **9**(76):2856–70. doi:10.1098/rsif.2012.0205
127. Wilkinson RW, Anderson G, Owen JJ, Jenkinson EJ. Positive selection of thymocytes involves sustained interactions with the thymic microenvironment. *J Immunol* (1995) **155**(11):5234–40.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 September 2013; paper pending published: 23 October 2013; accepted: 09 January 2014; published online: 04 February 2014.
Citation: Yates AJ (2014) Theories and quantification of thymic selection. Front. Immunol. ***5****:13. doi: 10.3389/fimmu.2014.00013*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Yates. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 February 2014
doi: 10.3389/fimmu.2014.00019

From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection

Maria Sawicka[1†], Gretta L. Stritesky[2†], Joseph Reynolds[1], Niloufar Abourashchi[1], Grant Lythe[1], Carmen Molina-París[1]* and Kristin A. Hogquist[2]*

[1] *Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds, UK*
[2] *Department of Laboratory Medicine and Pathology, Center for Immunology, University of Minnesota, Minneapolis, MN, USA*

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Ruy Ribeiro, Los Alamos National Laboratory, USA
Christian Schönbach, Nazarbayev University, Kazakhstan

Correspondence:
Carmen Molina-París, Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
e-mail: carmen@maths.leeds.ac.uk; Kristin A. Hogquist, Department of Laboratory Medicine and Pathology, Center for Immunology, University of Minnesota, 2101 6th Street SE, Minneapolis, MN 55414, USA
e-mail: hogqu001@umn.edu

[†] *Maria Sawicka and Gretta L. Stritesky have contributed equally to this work.*

Cells of the mature $\alpha\beta$ T cell repertoire arise from the development in the thymus of bone marrow precursors (thymocytes). $\alpha\beta$ T cell maturation is characterized by the expression of thousands of copies of identical $\alpha\beta$ T cell receptors and the CD4 and/or CD8 co-receptors on the surface of thymocytes. The maturation stages of a thymocyte are: (1) double negative (DN) (TCR^-, $CD4^-$ and $CD8^-$), (2) double positive (DP) (TCR^+, $CD4^+$ and $CD8^+$), and (3) single positive (SP) (TCR^+, $CD4^+$ or $CD8^+$). Thymic antigen presenting cells provide the appropriate micro-architecture for the maturation of thymocytes, which "sense" the signaling environment via their randomly generated TCRs. Thymic development is characterized by (i) an extremely low success rate, and (ii) the selection of a functional and self-tolerant T cell repertoire. In this paper, we combine recent experimental data and mathematical modeling to study the selection events that take place in the thymus after the DN stage. The stable steady state of the model for the pre-DP, post-DP, and SP populations is identified with the experimentally measured cell counts from 5.5- to 17-week-old mice. We make use of residence times in the cortex and the medulla for the different populations, as well as recently reported asymmetric death rates for CD4 and CD8 SP thymocytes. We estimate that 65.8% of pre-DP thymocytes undergo death by neglect. In the post-DP compartment, 91.7% undergo death by negative selection, 4.7% become CD4 SP, and 3.6% become CD8 SP. Death by negative selection in the medulla removes 8.6% of CD4 SP and 32.1% of CD8 SP thymocytes. Approximately 46.3% of CD4 SP and 27% of CD8 SP thymocytes divide before dying or exiting the thymus.

Keywords: thymocytes, negative selection, positive selection, death by neglect, mathematical model, steady state

1. INTRODUCTION

T cells are a major component of the adaptive immune system that play a crucial role in protection against a wide variety of pathogens. The T cell receptor (TCR) is generated by somatic recombination and has a vast potential to recognize foreign organisms. T cells do not recognize pathogens directly, but rather through binding pathogen fragments displayed by major histocompatibility complex (MHC) proteins on the surface of antigen presenting cells (APCs). Since MHC molecules are highly polymorphic, useful T cells must be selected for in each individual of the species. These T cells must have lineage specific effector functions that may include direct lysis, production of cytokines, and ability to regulate immune responses. Furthermore, some T cells have the potential to drive dangerous autoimmune responses (1). For all of these reasons, the development of a T cell repertoire is a highly specialized and tightly regulated process (2, 3). It takes place in a dedicated organ, the thymus, where unique properties of the microenvironment ensure the production of functional, yet self-tolerant T cells (4–6).

Multi-potent precursors travel from the bone marrow to the thymus through the blood. When they enter the thymus, the precursors that commit to the T cell lineage [or canonical early T cell progenitors (7)], after a 2 week period on average, transition from the double negative (DN) stage, where they do not express the co-receptors CD4 and CD8, to the double positive (DP) stage, where they express both co-receptors (6). At this stage, the majority of the cells have made productive TCR gene rearrangements and express a fully formed $\alpha\beta$ TCR on the cell surface. DP cells are located in the cortex region of the thymus, where they use their TCR to survey self-peptides presented by MHCs on cortical thymic epithelial cells (cTECs) (8). DPs that recognize self-peptide-MHC complexes with low affinity undergo positive selection, whereas those with high affinity are deleted by negative selection (3). Those DP that fail to recognize self-peptide-MHC will undergo apoptosis in a process referred to as death by neglect. The DP cells that are positively selected will then transition to the single positive (SP) stage, where they express either the CD4 or CD8 co-receptor, depending upon their MHC class specificity (9). MHC class specificity also dictates gene expression changes that will ultimately determine the effector functions of that T cell: generally, cytotoxicity for CD8 T cells and cytokine production for CD4 T cells. All positively selected cells, whether MHC class I or class II specific, up-regulate the chemokine receptor CCR7, which facilitates their migration to the medulla, where they undergo further selection events. The medulla contains medullary epithelial cells (mTECs) that express tissue-restricted antigens regulated by the nuclear factor Aire (10).

Exposure to tissue-restricted antigens allows for further deletion of T cells specific for self-antigens they may encounter in the periphery. Finally, those cells that have been positively selected, yet have avoided negative selection, will mature and migrate to the periphery (11).

Previous efforts to develop mathematical models of thymic selection have been based on deterministic approaches or cellular automata simulations. These studies have shown the importance of (i) thymic antigen diversity on the size of the selected T cell repertoire (12), (ii) death rates for the more differentiated thymocyte subsets (13), (iii) thymocyte proliferation and residence times (14), (iv) epithelial networks for thymocyte development and migration (15), (v) thymocyte competition for antigen (16), (vi) self-pMHC complexes expressed on dendritic cells (DCs) (17), (vii) receptor–ligand binding affinity (18), and (viii) a sharp threshold in TCR-ligand binding affinity that defines the boundary between negative and positive selection (19). Recent work by Ribeiro and Perelson (20) supports the need to develop appropriate mathematical models to interpret T cell receptor excision circles (or TREC) data, which are used to quantify thymic export (20). Sinclair et al. in Ref. (21) bring together experimental immunology with mathematical modeling to conclude that CD8 precursor thymocytes are more susceptible to death than CD4 precursors. This asymmetry in the death rates underlies the experimentally observed CD4:CD8 T cell ratio in the periphery.

Previous experimental studies have tried to determine the number of cells going through positive and negative selection in the thymus. However, reports estimating the relative number of cells undergoing negative selection compared to positive selection have been widely variable. Some find that very few cells undergo negative selection; others find that two times more cells undergo negative selection than positive selection (22–25). In this report, we develop a deterministic mathematical model of T cell development in the thymus. Some of us recently published a report where we used a novel approach ($Bim^{-/-}Nur77^{GFP}$ mice) that allowed us to calculate the number of cells undergoing positive and negative selection (26). Using previously published data on the relative life-span of DP and SP cells, we estimated the hourly rate of both positive and negative selection (26). In this manuscript, we make use of (i) a subset of this experimental data, and (ii) the asymmetric death rates observed for CD4 and CD8 precursor thymocytes (21), to develop two mathematical models that will enable us to estimate selection rates in the cortex and the medulla, and provide a quantitative measure for the stringency of thymic selection. The first model (see Section 2.1) allows the identification of the following parameters: DN thymocyte influx into the cortex, pre-DP and post-DP death rates, and pre-DP and post-DP differentiation rates. Under the assumption of asymmetric death rates for the CD4 and CD8 SP thymocytes (21), we extend the first model to provide estimates for the following medullary rates (see Section 2.2): CD4 and CD8 SP death, proliferation, and maturation rates.

2. MATERIALS AND METHODS

2.1. A FIRST MODEL OF THYMIC DEVELOPMENT AFTER THE DN STAGE

In this section, we introduce a deterministic model of thymocyte development after the DN stage. This first model will be required to calibrate the parameter values of the second model introduced in Section 2.2. In particular, and as described in Section 3.2, the first model allows the identification of parameter values for the following rates: ϕ, μ_1, μ_2, φ_1, and φ_2.

This mathematical model makes use of a data set obtained from the analysis of eight C57BL/6 wild type and Bim deficient mice (average age 9 weeks), that express a $Nur77^{GFP}$ transgene to indicate TCR signal strength experience (26). Flow cytometric analysis, as described in that study, used standard markers to define various stages of T cell development in the thymus. The $Nur77^{GFP}$ reporter and Bim deficiency were novel modifications that allowed us to quantify cells that normally would be deleted by strong TCR signaling. In the mathematical model, we consider the following thymocyte populations: n_1, the population of pre-selection DP thymocytes (double positive), that are $TCR\beta^{low}$ and $CD69^{low}$ (26), n_2, the population of post-selection DP thymocytes, that are $TCR\beta^{+}$ and $CD69^{high}$ (26), and n_3, the population of mature SP (single positive) thymocytes.

We assume that DN thymocytes differentiate to become pre-selection DP thymocytes with rate (cells/day) ϕ. We further assume that after the DN stage, thymocyte cell fate is determined by the TCR signal, which a given thymocyte has received. Sinclair et al. used CFSE labeling to show that there is no proliferation at the post-DP stage (see Figure A1 of their manuscript) (21). Stritesky et al. looked at proliferation in the post-DP pool with BrdU labeling, and found no evidence (26). We have, thus, only included proliferation in the SP thymocyte population (21, 26). The three populations, n_1, n_2, and n_3, are involved in the following selection events in the cortex and the medulla (see **Figure 1**):

- $\emptyset \xrightarrow{\phi} n_1$ – flux of DN thymocytes into compartment n_1,
- $n_1 \xrightarrow{\varphi_1} n_2$ – differentiation from pre-DP (n_1) to post-DP (n_2) thymocytes induced by TCR signal,
- $n_1 \xrightarrow{\mu_1} \emptyset$ – death by neglect of pre-DP thymocytes due to lack of (or weak) TCR signal,
- $n_2 \xrightarrow{\varphi_2} n_3$ – differentiation from post-DP (n_2) to SP (n_3) thymocytes sustained by intermediate TCR signal,
- $n_2 \xrightarrow{\mu_2} \emptyset$ – apoptosis of post-DP (n_2) thymocytes due to strong TCR signal,
- $n_3 \xrightarrow{\varphi_3}$ periphery – exit of SP thymocytes (n_3) to the periphery (thymic maturation),
- $n_3 \xrightarrow{\lambda_3} 2\,n_3$ – proliferation of SP thymocytes (n_3) in the medulla, and
- $n_3 \xrightarrow{\mu_3} \emptyset$ – apoptosis of SP (n_3) thymocytes due to strong TCR signal.

The time evolution of the three populations can be described by the following set of ordinary differential equations (ODEs), which are based on the selection events described above:

$$\begin{cases} \dfrac{dn_1}{dt} = \phi - \varphi_1 n_1 - \mu_1 n_1 \,, \\ \dfrac{dn_2}{dt} = \varphi_1 n_1 - \varphi_2 n_2 - \mu_2 n_2 \,, \\ \dfrac{dn_3}{dt} = \varphi_2 n_2 - \varphi_3 n_3 - \mu_3 n_3 + \lambda_3 n_3 \,. \end{cases} \tag{1}$$

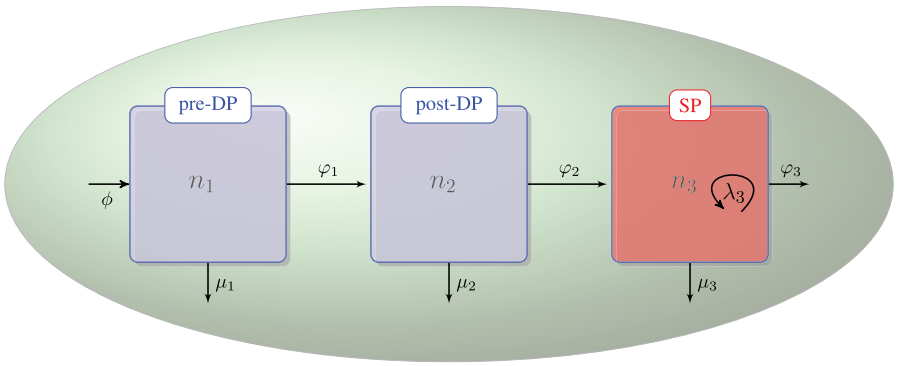


FIGURE 1 | Thymic development as hypothesized in the first model. The flux, ϕ, represents the differentiation of DNs into pre-DPs. Pre-DP thymocytes have two fates: further differentiation into the post-DP pool (φ_1) or death by neglect (μ_1). Post-DP thymocytes have two fates: further differentiation into the SP pool (φ_2) or death by apoptosis (μ_2). Finally, SP thymocytes have three fates: maturation and exit into the periphery (φ_3), death by apoptosis (μ_3), or proliferation (λ_3).

We are interested in studying the steady state of these populations, as the experimental data correspond to population cell numbers in the three stages (pre-DP, post-DP, and SP) for a steady state thymus (26). The steady state of the system of equations [equation (1)] is given by:

$$n_1^* = \frac{\phi}{\varphi_1 + \mu_1}\,, \quad n_2^* = \frac{n_1^* \varphi_1}{\varphi_2 + \mu_2}\,, \quad n_3^* = \frac{n_2^* \varphi_2}{\varphi_3 + \mu_3 - \lambda_3}\,. \quad (2)$$

This unique steady state exists if and only if $\varphi_3 + \mu_3 - \lambda_3 > 0$, so that we have $n_3^* > 0$. In order to study the linear stability of the steady state, we calculate A, the Jacobian matrix of equation (1), as follows:

$$A = \begin{pmatrix} -(\varphi_1 + \mu_1) & 0 & 0 \\ \varphi_1 & -(\varphi_2 + \mu_2) & 0 \\ 0 & \varphi_2 & -(\varphi_3 + \mu_3 - \lambda_3) \end{pmatrix}. \quad (3)$$

A is also the Jacobian matrix at the steady state $n^* = (n_1^*, n_2^*, n_3^*)$, as the system of ODEs [equation (1)] is linear. The three eigenvalues of A are given by (as the matrix is lower triangular):

$$\beta_1 = -(\varphi_1 + \mu_1)\,, \quad \beta_2 = -(\varphi_2 + \mu_2)\,, \quad \beta_3 = -(\varphi_3 + \mu_3 - \lambda_3)\,.$$

Therefore, the steady state [equation (2)] is stable, if and only if, $\varphi_3 + \mu_3 - \lambda_3 > 0$, which is also the condition for its existence. We conclude this section with the analytical solution of the system of ODEs [equation (1)], given initial conditions, which provides the time evolution of the three thymocyte populations:

$$\begin{aligned} n_1(t) &= n_1^* + n_1(0)\, e^{-(\varphi_1+\mu_1)t}\,, \\ n_2(t) &= n_2^* + \frac{n_1(0)\,\varphi_1}{[(\varphi_2+\mu_2) - (\varphi_1+\mu_1)]}\, e^{-(\varphi_1+\mu_1)t} \\ &\quad + n_2(0)\, e^{-(\varphi_2+\mu_2)t}\,, \end{aligned}$$

$$\begin{aligned} n_3(t) &= n_3^* + \frac{n_1(0)\,\varphi_1}{[(\varphi_2+\mu_2) - (\varphi_1+\mu_1)]} \\ &\quad \times \frac{\varphi_2}{[(\varphi_3+\mu_3-\lambda_3) - (\varphi_1+\mu_1)]}\, e^{-(\varphi_1+\mu_1)t} \\ &\quad + \frac{n_2(0)\,\varphi_3}{[(\varphi_3+\mu_3-\lambda_3) - (\varphi_2+\mu_2)]}\, e^{-(\varphi_2+\mu_2)t} \\ &\quad + n_3(0)\, e^{-(\varphi_3+\mu_3-\lambda_3)t}\,, \end{aligned} \quad (4)$$

where $n_1(0)$, $n_2(0)$, $n_3(0)$ represent the initial conditions for the thymocyte populations. Note that in the late time limit, that is, if $t \to +\infty$ and $\varphi_3 + \mu_3 - \lambda_3 > 0$, then $n_1(t) \to n_1^*$, $n_2(t) \to n_2^*$ and $n_3(t) \to n_3^*$, as it is the unique stable steady state.

2.2. A SECOND MODEL OF THYMIC DEVELOPMENT AFTER THE DN STAGE: CD4 AND CD8 SP THYMOCYTES

As described in Section 2.1, the first deterministic model will allow us to calibrate some of the parameters of a more comprehensive model, which we now introduce. We subdivide the SP thymocyte population in two classes: CD4 SP and CD8 SP thymocytes. This is an extension of the model introduced in the previous section, and is motivated by the fact that experimentally, SP thymocytes express either the CD4 or the CD8 co-receptor. We now have four different thymocyte populations to consider: n_1, the population of pre-selection DP (double positive) thymocytes, n_2, the population of post-selection DP thymocytes, n_4, the population of mature $CD4^+$ SP (single positive) thymocytes, and n_8, the population of mature $CD8^+$ SP (single positive) thymocytes.

As described in the previous section, we assume that DN thymocytes differentiate to become pre-selection DP thymocytes with rate (cells/day) ϕ, and that after the DN stage, thymocyte cell fate is determined by the TCR signal, which a given thymocyte has received. Thus, the four populations, n_1, n_2, n_4, and n_8, with $n_3 = n_4 + n_8$, are involved in the following selection events in the cortex and the medulla (see **Figure 2**):

- $\emptyset \xrightarrow{\phi} n_1$ – flux of DN thymocytes into compartment n_1,
- $n_1 \xrightarrow{\varphi_1} n_2$ – differentiation from pre-DP (n_1) to post-DP (n_2) thymocytes induced by TCR signal,

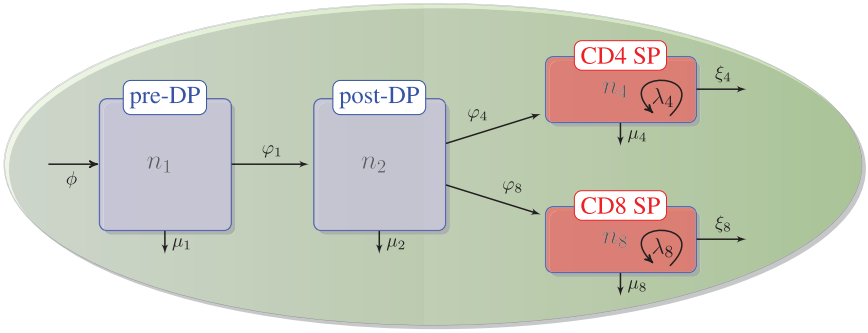


FIGURE 2 | Thymic development as hypothesized in the second model. We assume there is a flux, ϕ, that represents the differentiation of DNs into pre-DPs. Pre-DP thymocytes have two fates: further differentiation into the post-DP pool (φ_1) or death by neglect (μ_1). Post-DP thymocytes have three fates: further differentiation into the SP pool as CD4 SPs (φ_4), or CD8 SPs (φ_8), or death by apoptosis (μ_2). Finally, CD4 or CD8 SP thymocytes have three fates: maturation and exit into the periphery (ξ_4) or (ξ_8), death by apoptosis (μ_4) or (μ_8), or proliferation (λ_4) or (λ_8).

- $n_1 \xrightarrow{\mu_1} \emptyset$ – death by neglect of pre-DP thymocytes due to lack of (or weak) TCR signal,
- $n_2 \xrightarrow{\varphi_4} n_4$ – differentiation from post-DP (n_2) to CD4$^+$ SP (n_4) sustained by intermediate TCR signal,
- $n_2 \xrightarrow{\varphi_8} n_8$ – differentiation from post-DP (n_2) to CD8$^+$ SP (n_8) sustained by intermediate TCR signal,
- $n_2 \xrightarrow{\mu_2} \emptyset$ – apoptosis of post-DP (n_2) thymocytes due to strong TCR signal,
- $n_4 \xrightarrow{\xi_4}$ periphery – exit of CD4$^+$ SP thymocytes (n_4) to the periphery (thymic maturation),
- $n_8 \xrightarrow{\xi_8}$ periphery – exit of CD8$^+$ SP thymocytes (n_8) to the periphery (thymic maturation),
- $n_4 \xrightarrow{\lambda_4} 2\ n_4$ – proliferation of CD4$^+$ SP thymocytes (n_4) in the medulla,
- $n_8 \xrightarrow{\lambda_8} 2\ n_8$ – proliferation of CD8$^+$ SP thymocytes (n_8) in the medulla,
- $n_4 \xrightarrow{\mu_4} \emptyset$ – apoptosis of CD4$^+$SP (n_4) thymocytes due to strong TCR signal, and
- $n_8 \xrightarrow{\mu_8} \emptyset$ – apoptosis of CD8$^+$SP (n_8) thymocytes due to strong TCR signal.

We assume that all model parameters are positive, that is, μ_1, μ_2, μ_4, μ_8, φ_1, φ_2, φ_4, φ_8, ξ_4, ξ_8, ϕ, λ_4, $\lambda_8 > 0$, and note that the parameters and thymocyte populations of the first and second model are related by the following equations:

$$\varphi_2 = \varphi_4 + \varphi_8\,, \quad \xi_4\, n_4 + \xi_8\, n_8 = \varphi_3\, n_3\,,$$
$$\mu_4\, n_4 + \mu_8\, n_8 = \mu_3\, n_3\,, \quad \lambda_4\, n_4 + \lambda_8\, n_8 = \lambda_3\, n_3\,. \tag{5}$$

The time evolution of the four populations can be described by the following set of ODEs:

$$\begin{cases} \frac{dn_1}{dt} = \phi - \varphi_1 n_1 - \mu_1 n_1\,, \\ \frac{dn_2}{dt} = \varphi_1 n_1 - (\varphi_4 + \varphi_8) n_2 - \mu_2 n_2\,, \\ \frac{dn_4}{dt} = \varphi_4 n_2 - \xi_4 n_4 - \mu_4 n_4 + \lambda_4 n_4\,, \\ \frac{dn_8}{dt} = \varphi_8 n_2 - \xi_8 n_8 - \mu_8 n_8 + \lambda_8 n_8\,. \end{cases} \tag{6}$$

We are interested in studying the steady state of these populations, as the experimental data correspond to population cell numbers in the four stages (pre-DP, post-DP, CD4 SP, and CD8 SP) for a steady state thymus (26). The steady state of the system of equations [equation (6)] is given by:

$$n_1^* = \frac{\phi}{\varphi_1 + \mu_1}\,, \quad n_2^* = \frac{n_1^* \varphi_1}{\varphi_4 + \varphi_8 + \mu_2}\,,$$
$$n_4^* = \frac{n_2^* \varphi_4}{\xi_4 + \mu_4 - \lambda_4}\,, \quad n_8^* = \frac{n_2^* \varphi_8}{\xi_8 + \mu_8 - \lambda_8}\,. \tag{7}$$

This unique steady state exists if and only if $\xi_4 + \mu_4 - \lambda_4 > 0$ and $\xi_8 + \mu_8 - \lambda_8 > 0$, so that we guarantee $n_4^* > 0$ and $n_8^* > 0$. In order to study the linear stability of the steady state, we calculate B, the Jacobian matrix of equation (6), as follows:

$$B = \begin{pmatrix} -(\varphi_1 + \mu_1) & 0 & 0 & 0 \\ \varphi_1 & -(\varphi_4 + \varphi_8 + \mu_2) & 0 & 0 \\ 0 & \varphi_4 & -(\xi_4 + \mu_4 - \lambda_4) & 0 \\ 0 & \varphi_8 & 0 & -(\xi_8 + \mu_8 - \lambda_8) \end{pmatrix}. \tag{8}$$

B is also the Jacobian at the steady state $n^* = (n_1^*, n_2^*, n_4^*, n_8^*)$, as the system of ODEs is linear. The four eigenvalues of B are given by:

$$\beta_1 = -(\varphi_1 + \mu_1)\,, \quad \beta_2 = -(\varphi_4 + \varphi_8 + \mu_2)\,,$$
$$\beta_3 = -(\xi_4 + \mu_4 - \lambda_4)\,, \quad \beta_4 = -(\xi_8 + \mu_8 - \lambda_8)\,.$$

Therefore, the steady state equation (6) is stable if and only if $\xi_4 + \mu_4 - \lambda_4 > 0$ and $\xi_8 + \mu_8 - \lambda_8 > 0$, which is also the condition for its existence. We conclude this section with the analytical solution of the system of ODEs equation (6), given initial conditions, which provides the time evolution of the four thymocyte populations:

$$n_1(t) = n_1^* + n_1(0)\, e^{-(\varphi_1 + \mu_1)t}\,,$$
$$n_2(t) = n_2^* + \frac{n_1(0)\, \varphi_1}{[(\varphi_4 + \varphi_8 + \mu_2) - (\varphi_1 + \mu_1)]}\, e^{-(\varphi_1 + \mu_1)t} + n_2(0)\, e^{-(\varphi_4 + \varphi_8 + \mu_2)t}\,,$$

$$
\begin{aligned}
n_4(t) &= n_4^* + \frac{n_1(0)\,\varphi_1}{[(\varphi_4+\varphi_8+\mu_2)-(\varphi_1+\mu_1)]} \\
&\times \frac{\varphi_4+\varphi_8}{[(\xi_4+\mu_4-\lambda_4)-(\varphi_1+\mu_1)]}\, e^{-(\varphi_1+\mu_1)t} \\
&+ \frac{n_2(0)\,\xi_4}{[(\xi_4+\mu_4-\lambda_4)-(\varphi_4+\varphi_8+\mu_2)]}\, e^{-(\varphi_4+\varphi_8+\mu_2)t} \\
&+ n_4(0)\, e^{-(\xi_4+\mu_4-\lambda_4)t}\,, \\
n_8(t) &= n_8^* + \frac{n_1(0)\,\varphi_1}{[(\varphi_4+\varphi_8+\mu_2)-(\varphi_1+\mu_1)]} \\
&\times \frac{\varphi_4+\varphi_8}{[(\xi_8+\mu_8-\lambda_8)-(\varphi_1+\mu_1)]}\, e^{-(\varphi_1+\mu_1)t} \\
&+ \frac{n_2(0)\,\xi_8}{[(\xi_8+\mu_8-\lambda_8)-(\varphi_4+\varphi_8+\mu_2)]}\, e^{-(\varphi_4+\varphi_8+\mu_2)t} \\
&+ n_8(0)\, e^{-(\xi_8+\mu_8-\lambda_8)t}\,,
\end{aligned}
\tag{9}
$$

where $n_1(0), n_2(0), n_4(0), n_8(0)$ represent the initial conditions for the thymocyte populations. Note that in the late time limit, that is, if $t \to +\infty$ and $\xi_4+\mu_4-\lambda_4>0$ and $\xi_8+\mu_8-\lambda_8>0$, then $n_1(t) \to n_1^*, n_2(t) \to n_2^*, n_4(t) \to n_4^*$, and $n_8(t) \to n_8^*$, as it is the unique stable steady state.

3. RESULTS

3.1. PARAMETER ESTIMATION FOR THE FIRST MODEL (MEANS)

In this section, we make use of previously published experimental data (26) that provide thymocyte cell counts for the three subsets considered in the first model: pre-DPs, post-DPs, and SPs. The original experiments have been carried out for two types of mice: wild type mice ($\text{Bim}^{+/-}$) and Bim deficient mice ($\text{Bim}^{-/-}$). In this paper, we will only be considering the wild type experimental results. The data will allow us to provide experimental estimates for the steady state thymocyte cell counts: n_1^*, n_2^*, n_3^*. Note that, in order to estimate rates (with units of inverse time), thymocyte cell counts are not enough. Thus, we will make use of the additional knowledge provided by experimentally determined residence times for each population, τ_i, with $i=1,2,3$. If we make use of the model (see Section 2.1), the residence time in compartment i can be expressed as:

$$
\tau_i = \frac{1}{\varphi_i+\mu_i}\,, \quad \text{for} \quad i \in \{1,2,3\}\,.
$$

The experimental data (see **Table 1**) correspond to the number of cells (thymocytes) at steady state (26), in each of the thymic compartments considered in the mathematical models (see Sections 2.1 and 2.2), and for eight different mice ($j=1,2,\ldots,8$).

We have made use of the following average residence times in each compartment (27–29)

$$
\tau_1 = 60\text{ h} = 2.5\text{ days}\,, \quad \tau_2 = 16\text{ h} = 0.67\text{ days}\,,
$$
$$
\tau_3 = 96\text{ h} = 4\text{ days}\,.
$$

In order to derive estimates for the model parameters, we have carried out the following steps:

1. We make use of the experimentally determined mature SP thymocyte flux from the medulla to the periphery, which has been estimated to be $1\text{–}4\times10^6$ cells per day (14, 26, 30). This flux corresponds to about 1% of thymocytes leaving the thymus every day (30). Given this flux, which we denote by ϕ_{out}, n_3^*, and the fact that $\phi_{\text{out}} = \varphi_3\, n_3^*$, we can obtain an estimate for φ_3. We have chosen ϕ_{out} to be 2.5×10^6 cells per day (14, 30).
2. Given τ_3, φ_3, and the fact that $\tau_3 = \frac{1}{\mu_3+\varphi_3}$, we can obtain an estimate for μ_3.
3. Given τ_1, n_1^*, and the fact that $n_1^* = \phi\,\tau_1$, we can obtain an estimate for ϕ.
4. We also have $n_2^* = \varphi_1\,\tau_2\,n_1^*$, which, in principle, allows us to estimate φ_1. We make use of linear regression techniques to do so (31, 32).

Let us introduce a_1 by the following equation, $a_1 = \frac{n_2^*}{n_1^*}$, and make use of the experimental data to write: $n_2^{*,i} = a\, n_1^{*,i} + \epsilon^i$, for $i=1,2,\ldots,8$. Thus, the squared error is given by:

$$
E(a_1) = \sum_{i=1}^{8} \left(n_2^{*,i} - a_1\, n_1^{*,i}\right)^2 .
$$

We minimize $E(a_1)$ with respect to a_1, that is $\frac{dE}{da_1} = 0$. Solving for a_1, we obtain:

$$
a_1 = \frac{\sum_{i=1}^{8} n_1^{*,i}\, n_2^{*,i}}{\sum_{i=1}^{8} \left(n_1^{*,i}\right)^2}\,.
$$

Table 1 | Experimental steady state thymocyte cell counts for the wild type pre-DP, post-DP, CD4 SP, and CD8 SP populations.

Mouse	n_1^* (pre-DP) (cells)	n_2^* (post-DP) (cells)	n_3^* (SP) (cells)	n_4^* (SP CD4) (cells)	n_8^* (SP CD8) (cells)
1	82.58×10^6	9.30×10^6	18.36×10^6	13.85×10^6	4.51×10^6
2	142.19×10^6	19.94×10^6	26.20×10^6	18.73×10^6	7.46×10^6
3	89.00×10^6	5.98×10^6	15.98×10^6	11.88×10^6	4.10×10^6
4	29.32×10^6	2.09×10^6	5.61×10^6	4.40×10^6	1.21×10^6
5	29.32×10^6	2.09×10^6	5.61×10^6	4.40×10^6	1.21×10^6
6	51.26×10^6	5.93×10^6	9.01×10^6	6.85×10^6	2.16×10^6
7	64.48×10^6	6.81×10^6	11.64×10^6	9.03×10^6	2.61×10^6
8	218.94×10^6	15.42×10^6	40.20×10^6	29.46×10^6	10.74×10^6
Mean	$\mathbf{88.39}\times10^6$	$\mathbf{8.45}\times10^6$	$\mathbf{16.57}\times10^6$	$\mathbf{12.33}\times10^6$	$\mathbf{4.25}\times10^6$
Standard deviation	$\mathbf{60.11}\times10^6$	$\mathbf{5.89}\times10^6$	$\mathbf{11.05}\times10^6$	$\mathbf{7.94}\times10^6$	$\mathbf{3.12}\times10^6$

The bold font highlights the mean and the standard deviation from the individual mice data.

Given a_1, we can then estimate φ_1 from the equation $\varphi_1 = \frac{a_1}{\tau_2}$.

5. Given τ_1, φ_1, and the fact that $\tau_1 = \frac{1}{\mu_1+\varphi_1}$, we can obtain an estimate for μ_1.
6. We are now left with three remaining parameters: φ_2, μ_2, and λ_3. Given the experimental constraints on τ_1, τ_2, and τ_3, we assume that the average time to proliferate, $1/\lambda_3$, cannot be <7 days. Therefore, we consider λ_3 to be constrained in the interval $[1/7, \tau_3^{-1}]$, with time measured in days. We sample equally spaced values for λ_3, and for each value, we compute $\varphi_2 = \frac{n_3^*(\varphi_3+\mu_3-\lambda_3)}{n_2^*}$. The ratio $a_2 = \frac{n_3^*}{n_2^*}$ is computed using the linear regression method described above (see **Figure 3**). In this way, we obtain an estimate for φ_3. We note that the p-values for a_1 and a_2 are given by 7.57×10^{-3} and 6.85×10^{-3}, respectively (both smaller than the significance level $\alpha = 0.05$).
7. Given τ_2, φ_2, and the fact that $\tau_2 = \frac{1}{\mu_2+\varphi_2}$, we can obtain an estimate for μ_2.
8. From steps 6 and 7 above, we have generated (a table of) values for φ_2 and μ_2, given a fiducial value for λ_3 in the interval $[1/7, \tau_3^{-1}]$. The mice considered in the experimental study are 5.5–17 weeks old, and their thymus is in steady state (26). Thus, we expect that the parameter values can only be accepted if the corresponding system of ODEs attains steady state by 3 weeks. Therefore, we only accept parameter values which provide thymocyte cell counts at time $t = 21$ days that are within one standard deviation from the experimentally determined values (see **Table 1**). That is, we impose for the given parameter set that the mathematically predicted value $n_i(t = 21 \text{ days})$ belongs to the interval $n_i^* \pm \sigma_i$, with $i = 1, 2, 3$, and where n_i^* is the (experimental) mean number of cells in compartment i, and σ_i is the (experimental) standard deviation in compartment i, as given in **Table 1**.

We obtain the following parameter values:

$$\phi = 35.350 \times 10^6 \text{ cells/day}\,, \quad \mu_1 = 0.263 \text{ day}^{-1}\,,$$
$$\mu_3 = 0.099 \text{ day}^{-1}\,, \quad \varphi_1 = 0.137 \text{ day}^{-1}\,, \quad \varphi_3 = 0.151 \text{ day}^{-1}\,,$$

and

$$\mu_2 \in [1.295, 1.443] \text{ day}^{-1}\,, \quad \varphi_2 \in [0.050, 0.198] \text{ day}^{-1}\,,$$
$$\lambda_3 \in [0.143, 0.250] \text{ day}^{-1}\,.$$

These parameters imply the following thymic selection rates:

3.1.1. Death rates

9.7×10^5 cells/h die by neglect in compartment 1 ($\mu_1\ n_1^*$), 4.8×10^5 cells/h die by negative selection in compartment 2 ($\mu_2\ n_2^*$), and 6.9×10^4 cells/h die by negative selection in compartment 3 ($\mu_3\ n_3^*$).

3.1.2. Differentiation rates

5.0×10^5 cells/h are positively selected in compartment 1, that is, become post-DP from pre-DP ($\varphi_1\ n_1^*$), 4.4×10^4 cells/h are positively selected in compartment 2, that is, become SP from post-DP ($\varphi_2\ n_2^*$), and 1.0×10^5 cells/h leave compartment 3 to go to the periphery ($\varphi_3\ n_3^*$).

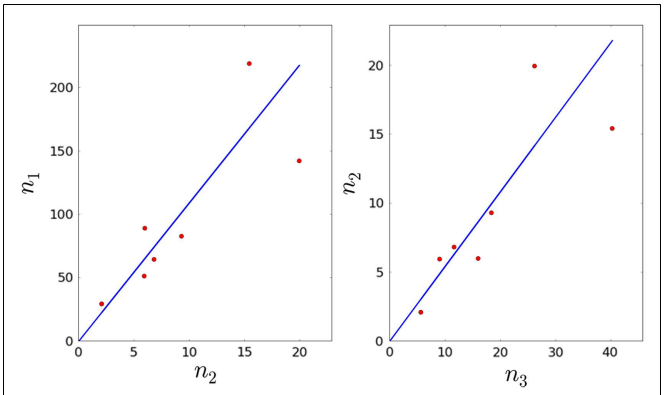


FIGURE 3 | Linear regression plots for the first model.

3.1.3. Proliferation rate

1.3×10^5 cells/h proliferate in compartment 3 ($\lambda_3\ n_3^*$).

We have also computed the stringency of thymic selection, which we define as given by the ratio:

$$\frac{\varphi_3\ n_3^*}{\phi} = 6.79\%\,.$$

Finally, we have computed the (per cell) probability to die, given that the cell is in compartment i ($i = 1, 2, 3$), as well as the (per cell) probability to proliferate in the medulla. We have obtained:

$$p_1 = \frac{\mu_1}{\mu_1+\varphi_1} = 65.8\%\,, \quad p_2 = \frac{\mu_2}{\mu_2+\varphi_2} = 91.7\%\,,$$
$$p_3 = \frac{\mu_3}{\mu_3+\varphi_3+\lambda_3} = 22.9\%\,, \quad q_3 = \frac{\lambda_3}{\mu_3+\varphi_3+\lambda_3} = 42.2\%\,.$$

3.2. PARAMETER ESTIMATION FOR THE SECOND MODEL (MEANS)

In this section, we make use of previously published experimental data (26) that provide thymocyte cell counts for the four subsets considered: pre-DPs, post-DPs, CD4 SPs, and CD8 SPs. We only make use of the experimental data for the wild type mice. The data will allow us to provide experimental estimates for the steady state thymocyte cell counts: $n_1^*, n_2^*, n_4^*, n_8^*$. As described in Section 3.1, we also need residence times for each population subset, τ_i, with $i = 1, 2, 4, 8$. If we make use of the model (see Section 2.2), the residence time in compartment i can be expressed as:

$$\tau_i = \frac{1}{\varphi_i + \mu_i}\,, \quad \text{for} \quad i \in \{1, 2\}\,, \quad \text{and}$$
$$\tau_i = \frac{1}{\xi_i + \mu_i}\,, \quad \text{for} \quad i \in \{4, 8\}\,.$$

Recent experimental data provide support for asymmetric death rates in the CD4 and CD8 SP compartments (21). The estimated death rates for CD4 and CD8 SP thymocytes are[1] $\mu_4 = 0.04 \text{ day}^{-1}$ and $\mu_8 = 0.11 \text{ day}^{-1}$. We also make use of the estimates derived in Section 3.1 for ϕ, μ_1, μ_2, φ_1, and φ_2. Finally,

[1] Private communication from Ben Seddon and Andy Yates.

the average residence times in each compartment, as described in Section 3.1, are given by:

$$\tau_1 = 60\text{ h} = 2.5\text{ days}\,, \quad \tau_2 = 16\text{ h} = 0.67\text{ days}\,,$$
$$\tau_4 = 96\text{ h} = 4\text{ days}\,, \quad \tau_8 = 96\text{ h} = 4\text{ days}\,.$$

In order to derive estimates for the model parameters, we have carried out the following steps:

1. Given τ_4, μ_4, and the fact that $\tau_4 = \frac{1}{\mu_4+\xi_4}$, we can obtain an estimate for ξ_4.
2. In the same way, given τ_8, μ_8, and the fact that $\tau_8 = \frac{1}{\mu_8+\xi_8}$, we can obtain an estimate for ξ_8.
3. We are now left with four remaining parameters: $\varphi_4, \varphi_8, \lambda_4$, and λ_8. We know that $\varphi_2 = \varphi_4 + \varphi_8$. We sample φ_4 in the interval $[0, \varphi_2]$, where φ_2 is the mean value of the interval obtained in Section 2.1, and for each fiducial value for φ_4, we compute the corresponding value for φ_8.
4. Given τ_4, φ_4, and the fact that $n_4^* = \frac{n_2^*\varphi_4}{\tau_4^{-1}-\lambda_4}$, we can compute the fraction $a_3 = \frac{n_2^*}{n_4^*}$ by linear regression (see **Figure 4**), and thus obtain an estimate for λ_4. Note that we will reject values of λ_4 that imply the proliferation time is larger than 7 days (see Section 3.1).
5. In Section 3.1, we obtained an estimate for the mean of λ_3, and we know that $\lambda_4\ n_4^* + \lambda_8\ n_8^* = \lambda_3\ n_3^*$. As before, we can compute the fractions $a_4 = \frac{n_4^*}{n_8^*}$ and $a_5 = \frac{n_3^*}{n_8^*}$ by linear regression (see **Figure 4**), and thus obtain an estimate for λ_8. Note that we will reject values of λ_8 that imply the proliferation time is larger than 7 days (see Section 3.1). We note that the p-values for a_3, a_4, and a_5 are given by 8.43×10^{-3}, 3.33×10^{-7}, and 4.56×10^{-8}, respectively (smaller than the significance level).
6. From steps 3, 4, and 5 above, we have generated (a table of) values for φ_8, λ_4, and λ_8, given a fiducial value for φ_4 in the interval $[0, \varphi_2]$. As discussed in Section 3.1, we only accept parameter values which provide thymocyte cell counts at time $t = 21$ days that are within one standard deviation from the experimentally determined values (see **Table 1**).

We obtain the following parameter values:

$$\mu_4 = 0.04\text{ day}^{-1}\,, \quad \mu_8 = 0.11\text{ day}^{-1}\,,$$
$$\xi_4 = 0.21\text{ day}^{-1}\,, \quad \xi_8 = 0.14\text{ day}^{-1}\,,$$

and

$$\varphi_4 = 0.070\text{ day}^{-1}\,, \quad \varphi_8 = 0.054\text{ day}^{-1}\,,$$
$$\lambda_4 = 0.216\text{ day}^{-1}\,, \quad \lambda_8 = 0.093\text{ day}^{-1}\,.$$

These parameters imply the following thymic selection rates:

3.2.1. Death rates

2.05×10^4 cells/h die by negative selection in compartment 4 ($\mu_4\ n_4^*$) and 1.95×10^4 cells/h die by negative selection in compartment 8 ($\mu_8\ n_8^*$).

3.2.2. Differentiation rates

2.50×10^4 cells/h are CD4 positively selected in compartment 2, that is, become CD4 SP from post-DP ($\varphi_4\ n_2^*$), 1.90×10^4 cells/h are CD8 positively selected in compartment 2, that is, become CD8 SP from post-DP ($\varphi_8\ n_2^*$), 1.08×10^5 cells/h leave compartment 4 to go to the periphery ($\xi_4\ n_4^*$), and 2.48×10^4 cells/h leave compartment 8 to go to the periphery ($\xi_8\ n_8^*$).

3.2.3. Proliferation rates

11.10×10^4 cells/h proliferate in compartment 4 ($\lambda_4\ n_4^*$) and 1.60×10^4 cells/h proliferate in compartment 8 ($\lambda_8\ n_8^*$).

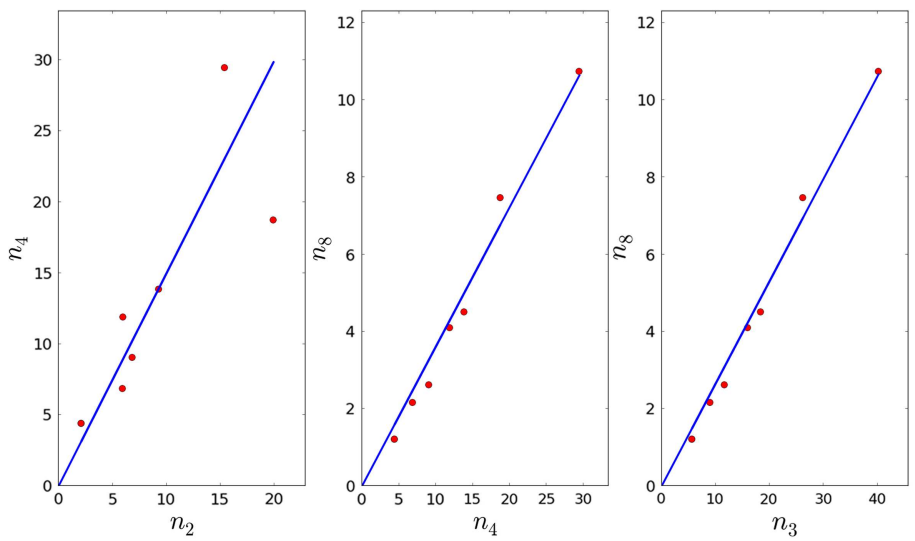


FIGURE 4 | Linear regression plots for the second model.

We can compute the stringency of thymic selection, defined by the ratio:

$$\frac{\xi_4\, n_4^* + \xi_8\, n_8^*}{\phi} = 8.96\% \,.$$

We can provide an estimate for the cortical positive selection probabilities, that is the (per post-DP cell) probability to become a CD4 SP or a CD8 SP, and the probability to be negatively selected in the cortex. We have obtained:

$$s_4 = \frac{\varphi_4}{\mu_2 + \varphi_4 + \varphi_8} = 4.7\% \,, \quad s_8 = \frac{\varphi_8}{\mu_2 + \varphi_4 + \varphi_8} = 3.6\% \,,$$
$$p_2 = \frac{\mu_2}{\mu_2 + \varphi_4 + \varphi_8} = 91.7\% \,.$$

Finally, we have computed the (per cell) probability to die, given that the cell is in compartment i, as well as the (per cell) probability to proliferate in the medulla. We obtain:

$$p_4 = \frac{\mu_4}{\mu_4 + \xi_4 + \lambda_4} = 8.6\% \,, \quad q_4 = \frac{\lambda_4}{\mu_4 + \xi_4 + \lambda_4} = 46.3\% \,,$$
$$p_8 = \frac{\mu_8}{\mu_8 + \xi_8 + \lambda_8} = 32.1\% \,, \quad q_8 = \frac{\lambda_8}{\mu_8 + \xi_8 + \lambda_8} = 27.0\% \,.$$

These probabilities imply that the probability to exit the thymus as a mature CD4 thymocyte (that has already reached the medulla) is given by $\frac{100-(8.6+46.3)}{100}$, which is 45.1%, and the probability to exit as a mature CD8 thymocyte (that has already reached the medulla) is given by $\frac{100-(32.1+27.0)}{100}$, which is 40.9%.

3.3. SENSITIVITY ANALYSIS

In this section, we explore the sensitivity of the parameters to perturbations in the experimental data. For the first model, the experimental data are given in terms of the following eight quantities:

$$\theta = (\tau_1, \tau_2, \tau_3, \phi_{\text{out}}, a_1, a_2, \bar{n}_3, \bar{n}_1) \,,$$

where a_1, a_2 are the coefficients of the linear regression of $\frac{n_2^*}{n_1^*}$ and $\frac{n_3^*}{n_2^*}$, respectively, and $\bar{n}_i$ is the experimental mean value of n_i.

We perturb each entry of the vector θ by adding and subtracting 10% of its value. Therefore, we now have two values for θ_i, equal to $\theta_i + \frac{1}{10}\theta_i$ and $\theta_i - \frac{1}{10}\theta_i$. Consequently, we have 2^8 sets of θ, which will be used to compute the corresponding model parameters as described in Section 3.1. Parameter values will only be accepted if they provide a stable solution before $t = 21$ days.

For the second model, the experimental data is given in terms of the following seven quantities:

$$\theta = (\tau_4, \tau_8, \mu_4, \mu_8, a_3, a_4, a_5) \,,$$

where a_3, a_4, a_5 are the coefficients of the linear regression of $\frac{n_2^*}{n_4^*}$, $\frac{n_4^*}{n_8^*}$, and $\frac{n_3^*}{n_8^*}$, respectively. We have made use of the means of the following parameters of the first model: ϕ, φ_1, μ_1, φ_2, μ_2, λ_3.

Table 2 | Means, 95% trimmed and minimum–maximum intervals of the model parameters.

Parameter	Mean value	95% Trimmed interval	Minimum–maximum interval range
ϕ	35.86×10^6 cells/day	(35.65×10^6, 35.07×10^6) cells/day	(28.93, 43.21 $\times 10^6$) cells/day
φ_1	0.139 day^{-1}	(0.138, 0.140) day^{-1}	(0.112, 0.167) day^{-1}
φ_2	0.136 day^{-1}	(0.134, 0.139) day^{-1}	(0.041, 0.274) day^{-1}
φ_4	0.140 day^{-1}	(0.136, 0.145) day^{-1}	(0.060, 0.264) day^{-1}
φ_8	0.134 day^{-1}	(0.129, 0.138) day^{-1}	(0.010, 0.214) day^{-1}
μ_1	0.265 day^{-1}	(0.263, 0.267) day^{-1}	(0.196, 0.333) day^{-1}
μ_2	1.372 day^{-1}	(1.365, 1.378) day^{-1}	(1.083, 1.618) day^{-1}
μ_4	0.040 day^{-1}	*n/a*	(0.036, 0.044) day^{-1}
μ_8	0.110 day^{-1}	*n/a*	(0.099, 0.121) day^{-1}
λ_4	0.181 day^{-1}	(0.179, 0.184) day^{-1}	(0.116, 0.226) day^{-1}
λ_8	0.085 day^{-1}	(0.080, 0.090) day^{-1}	(0.078, 0.092) day^{-1}
ξ_4	0.231 day^{-1}	(0.230, 0.233) day^{-1}	(0.229, 0.233) day^{-1}
ξ_8	0.152 day^{-1}	(0.150, 0.154) day^{-1}	(0.149, 0.155) day^{-1}

We perturb each entry of the vector θ as described above. Therefore, we have $2^7 n_{\varphi_4}$ sets of θ, with n_{φ_4}, the number of different values considered for φ_4 in the interval $(0, \varphi_2)$. Parameter values will only be accepted if they provide a stable solution before $t = 21$ days.

The results of the sensitivity analysis, with 95% trimmed intervals[2] and minimum–maximum interval ranges, are given in **Table 2**.

3.4. VARIABILITY IN THE SELECTION RATES

The (trimmed and minimum–maximum) intervals derived in Section 3.3 allow us to estimate the variability in the different selection rates discussed in Sections 3.1 and 3.2. For example, given variations in the parameters, the corresponding variations in the selection rates can be shown to be:

$$\Delta p_i = \frac{1}{(\mu_i + \varphi_i)^2}(\varphi_i \Delta\mu_i + \mu_i \Delta\varphi_i) \quad \text{for} \quad i = 1, 2 \,, \tag{10}$$

$$\Delta p_3 = \frac{1}{(\mu_3 + \varphi_3 + \lambda_3)^2}[(\varphi_3 + \lambda_3)\Delta\mu_3 + \mu_3 \Delta\varphi_3 + \mu_3 \Delta\lambda_3] \,, \tag{11}$$

$$\Delta q_3 = \frac{1}{(\mu_3 + \varphi_3 + \lambda_3)^2}[\lambda_3 \Delta\mu_3 + \lambda_3 \Delta\varphi_3 + (\mu_3 + \varphi_3)\Delta\lambda_3] \,, \tag{12}$$

$$\Delta s_i = \frac{1}{(\mu_2 + \varphi_i + \varphi_j)^2}[\varphi_i \Delta\mu_2 + \varphi_i \Delta\varphi_j + (\mu_2 + \varphi_j)\Delta\varphi_i] \quad \text{for} \quad i = 4, j = 8 \quad \text{or} \quad i = 8, j = 4 \,, \tag{13}$$

$$\Delta p_i = \frac{1}{(\mu_i + \xi_i + \lambda_i)^2}[\mu_i \Delta\xi_i + \mu_i \Delta\lambda_i + (\xi_i + \lambda_i)\Delta\mu_i] \quad \text{for} \quad i = 4, 8 \,, \tag{14}$$

[2]We define the 95% trimmed interval to be the result of the sensitivity analysis after trimming the lower and upper 2.5% of values.

$$\Delta q_i = \frac{1}{(\mu_i + \xi_i + \lambda_i)^2}[\lambda_i \Delta\xi_i + \lambda_i \Delta\mu_i + (\xi_i + \mu_i)\Delta\lambda_i] \quad \text{for} \quad i = 4, 8\,. \tag{15}$$

We present in **Table 3** the variability of the selection rates.

4. DISCUSSION

We have brought together experimental data with a mathematical compartment model [similar to other progression models of CD4 and CD8 T cell development (13, 14, 18, 21, 33)] to provide estimates for the selection events that take place in the thymus. We have made use of a range of experimental data: (i) steady state thymocyte cell counts (26), mean residence times in each compartment (27–29), murine thymic export rate (14, 26, 30), and recently reported asymmetric death rates for the CD4 SP and CD8 SP thymocytes (21). Our preliminary results support the unexpectedly high death rate in the post-DP thymocyte population observed in Ref. (21). We note that our approach is unrelated to that of Sinclair et al. both experimentally and mathematically (21). This rate, μ_2, has been estimated to be at least an order of magnitude larger than any of the other death rates in the pre-DP, CD4 SP, or CD8 SP pools (see **Table 2**). In terms of selection rates, our analysis yields the following: pre-selection thymocytes (pre-DPs) have a 65.8% probability of dying by neglect in the cortex, and a 34.2% probability of becoming post-selection thymocytes (post-DPs). At the post-selection stage, post-DPs have a 91.7% probability of dying by negative selection (apoptosis) in the cortex, a 4.7% probability of becoming CD4 SPs, and a 3.6% probability of becoming CD8 SPs. In the medulla, CD4 SPs have an 8.6% probability of dying by negative selection (apoptosis), whereas CD8 SPs have a 32.1% probability of dying by negative selection. CD4 SPs have a 45.1% probability of exiting the thymus and reaching the periphery as mature thymocytes, whereas that probability for CD8 SPs is only 40.9%. Finally, the data supports some level of cellular proliferation in the medulla, with CD4 SPs having a 46.3% probability of proliferation and CD8 SPs a 27% probability.

Earlier work by Mehr and collaborators combined experimental and theoretical approaches to estimate thymic selection rates (13, 33), neglected death rates in the medulla, but considered potential feedback from mature T cells. In agreement with these authors, our results indicate that thymocyte death is highest at the post-DP stage. However, as death in the medulla had been neglected, these authors concluded that the CD4:CD8 ratio in SP thymocytes is determined by the differentiation rates. In this paper, we have made use of CD4 and CD8, or medullary, death rates, which allowed us to directly compare cortical (DP) to medullary (SP) death rates. Furthermore, our approach allowed us to conclude that medullary, or SP, death was due to negative selection, as it was rescued by Bim deficiency (26). Sinclair et al. also recently addressed the temporal dynamics of thymic selection using an unrelated approach (both experimentally and mathematically) (21). While their experimental approach did not allow them to distinguish death by negative selection from death by other mechanisms, their overall finding was consistent with ours, that thymocyte death is highest at the post-DP stage.

Table 3 | Selection rate values (initial and after perturbation) and their variability intervals.

Rate	Initial value (%)	After perturbation (%)	Δ Value (%)	Δ Min–max (%)
p_1	65.8	65.66	±0.76	±20.55
p_2	91.7	90.98	±0.49	±17.28
p_3	22.91	24.07	±0.90	±28.22
q_3	42.24	41.74	±0.99	±30.53
s_4	4.69	8.51	±0.80	±15.16
s_8	3.61	8.13	±0.79	±15.03
p_4	8.59	8.84	±0.37	±4.91
q_4	46.29	40.07	±1.29	±20.49
p_8	32.12	31.71	±2.25	±35.37
q_8	27.0	24.5	±3.58	±64.81

Further attempts to quantify thymic selection rates making use of mathematical models also include those of Faro et al. (12). The mathematical model developed by Faro and collaborators does not include time dynamics, but describes the relationship between the number of selecting ligands and the probability of selection of a given thymocyte. Thomas-Vaslin et al. (14) obtained estimates of thymic selection rates, using an experimental procedure that temporarily blocks thymic output and a mathematical model in which rates of transit from compartment to compartment depend on the number of cell divisions. Their model can capture the thymic "conveyor belt" (34, 35) scheme, but requires more differential equations and more parameters than equation (6). Despite the differences between their theoretical and experimental models and ours, similar estimates for thymic selection rates are found. For example, we estimate that 1.2 million post-DP become CD4 SP thymocytes per day and 0.5 million post-DP become CD8 SP thymocytes per day; their estimates are 0.9 and 0.2 million, respectively. Finally, we estimate that 2.6 million CD4 SP thymocytes per day and 0.6 million CD8 SP thymocytes per day exit the thymus. Their estimates are 2.4 and 0.5 million, respectively.

Our estimates of how many CD4 and CD8 SP thymocytes survive and exit the thymus reflect the skewed CD4:CD8 SP thymocyte ratio observed in C57BL/6 mice, which is approximately 4:1 (36). This ratio is similar to the reported CD4:CD8 ratio of recent thymic emigrants (37), and raises the question of what accounts for the CD4 bias. While we were able to determine death and differentiation rates for both CD4 and CD8 SP thymocytes (see **Table 2**), our experimental approach did not allow us to determine what fraction of the post-DP pool was MHC class I versus II restricted. Therefore, we could not address the issue of when and how the CD4:CD8 bias becomes established. The approach of Sinclair et al., which used MHC class I and class II deficiency, allowed them to address this question. Their data suggest that the skewed CD4:CD8 ratio reflects asymmetry in post-selection DP death rates, rather than more efficient positive selection of CD4 compared to CD8 thymocytes (21). Yet, the parameter estimation allows us to compare the following different CD4:CD8 ratios (see

Section 3.2): (i) the CD4:CD8 ratio of positive selection in the post-DP pool (differentiation from post-DP to either CD4 SP or CD8 SP) is given by $\frac{\varphi_4}{\varphi_8} \approx 5:4$, (ii) the CD4:CD8 ratio in the SP pool is given by $\frac{n_4^*}{n_8^*} \approx 3:1$, and (iii) the CD4:CD8 ratio of positive selection in the SP pool (differentiation from SP to peripheral early thymic emigrants) is given by $\frac{\xi_4 n_4^*}{\xi_8 n_8^*} \approx 4:1$. Our observations indicate that the CD4 bias is progressively established, as the thymocytes mature from the post-DP stage until the exit of the SP stage to migrate to the periphery.

Our mathematical analysis has also allowed us to estimate the stringency of thymic selection, defined by:

$$\sigma = \frac{\xi_4\, n_4^* + \xi_8\, n_8^*}{\phi} = 8.96\%\,,$$

that is, the ratio between the number of thymocytes per unit time that exit the thymus and the number of thymocytes per unit time that enter the pre-DP stage. The sensitivity analysis described in Section 3.3 allows us to provide a value of $\Delta\sigma = 0.2\%$, where we have made use of the minimum–maximum interval ranges (see fourth column of **Table 2**). A different measure of stringency could be based on the probability of a cell surviving the maturation process. In our notation, this would correspond to the following:

$$(1-p_1)\times(1-p_2)\times(1-p_3) = 2.19\%\,.$$

We note that this measure of stringency is the probability of not dying in any of the three compartments considered in the model (pre-DP, post-DP, and SP). As discussed in Appendix A.1, and given that in the SP pool, thymocytes may proliferate, there is a need to consider this special case. Our estimates suggest that a population of 10^3 pre-DP thymocytes will yield 69 CD4 and 25 CD8 SP thymocytes that leave the medulla to get incorporated into the peripheral naive T cell pool (see details in Appendix A.1).

The sensitivity analysis (see Section 3.3) and the variability of the selection rates derived from it (see Section 3.4) give us the confidence to conclude, that our parameter estimation is robust. We are aware that the experimental data we have made use of [steady state thymocyte cell counts (26)] do not provide the exquisite time resolution described in Ref. (21). However, the supporting mathematical model described in Section 2.2, allows us to obtain the time evolution of the thymocyte populations, once the parameters have been estimated. In **Figure 5**, we plot the time evolution of the total number of cells in each compartment of the mathematical model: pre-DP, post-DP, CD4 SP, and CD8 SP thymocytes. We start with no cells at time zero, $n_i(t=0)=0$ for $i=1, 2, 4, 8$. Trajectories have been plotted for a period of 6 weeks and have been computed for every permutation of the parameter set presented in **Table 2**. The subset of parameters shared with the simple model ($\phi, \varphi_1, \mu_1, \mu_2$), were fixed at their mean values. Thus, 548 distinct parameter sets were generated. The system of equations (6) was solved using a fourth order Runge–Kutta method (Python source code).

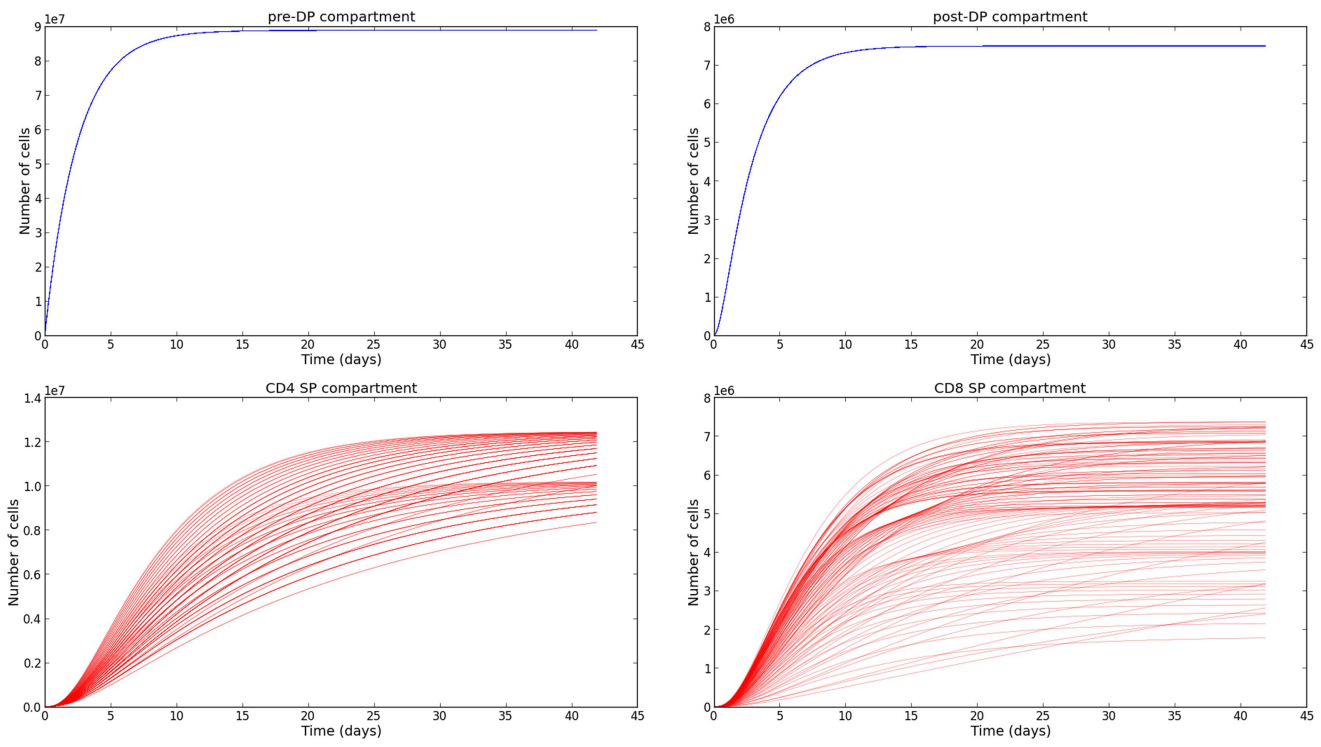


FIGURE 5 | Time evolution of the thymocyte populations in the second model. The different trajectories correspond to the parameter values and ranges described in **Table 2**.

The approaches introduced in this paper have shed some light on the probabilities and timescales that characterize cellular fate in the thymus after the DN stage. We plan to generalize the mathematical model introduced here, making use of experimental data for the strength of TCR binding in Nur77^{GFP} mice (26), to investigate issues such as the death rate in the post-DP pool and the CD4:CD8 ratio. Our model assumes that all progenitors in a particular pool behave with identical kinetics, i.e., move through the various stages of selection at the same rate. Future model refinements will come from consideration of the heterogeneity of the pools, which are known to include cells that will become iNKT cells, regulatory T cells, and intra-epithelial lymphocytes (2). It is also possible that progenitors of the same general class move through the selection process with different kinetics (34). The models introduced here can serve as a first step to study human thymic selection, although comprehensive data on human thymic subsets, their sizes, and residence times are not yet available. It would be of great interest to apply the model to data on thymic subsets and cellularity in children, keeping in mind that residence times of human subsets may differ from murine ones (38). Finally, we note that we have not mentioned the relevance of cytokines, such as IL-7, during thymic development. Some differences have already been described for the role of IL-7R in human versus mouse T cell development (38, 39). We hope in the near future to combine mechanistic mathematical models of IL-7 and IL-7R (40) with the T cell development model introduced here to address these issues.

ACKNOWLEDGMENTS

We thank Ed Palmer, Robin Callard, Andy Yates, and Ben Seddon for helpful discussions, and Andy Yates and Ben Seddon for allowing us to make use of their mathematical estimates for μ_4 and μ_8. Grant Lythe and Carmen Molina-París thank Bill Hlavacek, Ruy Ribeiro, and Alan Perelson (Los Alamos National Laboratory) for their hospitality. This work was supported by a BBSRC Research Development Fellowship BB/G023395/1 (to Carmen Molina-París), an FP7 IRSES INDOEUROPEAN-MATHDS Network PIRSES-GA-2012-317893 (to Grant Lythe and Carmen Molina-París), and National Institutes of Health Grants R37 AI39560 (to Kristin A. Hogquist) and F32 AI100346 (to Gretta L. Stritesky). Maria Sawicka, Joseph Reynolds, Niloufar Abourashchi, Grant Lythe, and Carmen Molina-París acknowledge the hospitality of the Max Planck Institute for the Physics of Complex Systems, Dresden, and the International Centre for Mathematical Sciences, Edinburgh, where part of this work was discussed and presented.

REFERENCES

1. Anderson G, Lane P, Jenkinson E. Generating intrathymic microenvironments to establish T-cell tolerance. *Nat Rev Immunol* (2007) **7**(12):954–63. doi:10.1038/nri2187
2. Stritesky GL, Jameson SC, Hogquist KA. Selection of self-reactive T cells in the thymus. *Annu Rev Immunol* (2012) **30**:95. doi:10.1146/annurev-immunol-020711-075035
3. Palmer E. Negative selection: clearing out the bad apples from the T-cell repertoire. *Nat Rev Immunol* (2003) **3**(5):383–91. doi:10.1038/nri1085
4. Jameson S, Hogquist K, Bevan M. Positive selection of thymocytes. *Annu Rev Immunol* (1995) **13**(1):93–126. doi:10.1146/annurev.iy.13.040195.000521
5. Werlen G, Hausmann B, Naeher D, Palmer E. Signaling life and death in the thymus: timing is everything. *Science* (2003) **299**(5614):1859. doi:10.1126/science.1067833
6. Petrie H, Zúñiga-Pflücker J. Zoned out: functional mapping of stromal signaling microenvironments in the thymus. *Immunology* (2007) **25**(1):649. doi:10.1146/annurev.immunol.23.021704.115715
7. Zúñiga-Pflücker JC. When three negatives made a positive influence in defining four early steps in T cell development. *J Immunol* (2012) **189**(9):4201–2. doi:10.4049/jimmunol.1202553
8. Takada K, Ohigashi I, Kasai M, Nakase H, Takahama Y. Development and function of cortical thymic epithelial cells. *Curr Top Microbiol Immunol* (2013) **373**:1–17.
9. Singer A, Adoro S, Park J. Lineage fate and intense debate: myths, models and mechanisms of CD4-versus CD8-lineage choice. *Nat Rev Immunol* (2008) **8**(10):788–801. doi:10.1038/nri2416
10. Anderson MS, Su MA. Aire and T cell development. *Curr Opin Immunol* (2011) **23**(2):198–206. doi:10.1016/j.coi.2010.11.007
11. Klein L. Dead man walking: how thymocytes scan the medulla. *Nat Immunol* (2009) **10**(8):809–11. doi:10.1038/ni0809-809
12. Faro J, Velasco S, Gonzalez-Fernandez A, Bandeira A. The impact of thymic antigen diversity on the size of the selected T cell repertoire. *J Immunol* (2004) **172**(4):2247.
13. Mehr R, Globerson A, Perelson A. Modeling positive and negative selection and differentiation processes in the thymus. *J Theor Biol* (1995) **175**(1):103–26. doi:10.1006/jtbi.1995.0124
14. Thomas-Vaslin V, Altes H, de Boer R, Klatzmann D. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *J Immunol* (2008) **180**(4):2240.
15. Souza-e Silva H, Savino W, Feijóo R, Vasconcelos A. A cellular automata-based mathematical model for thymocyte development. *PLoS One* (2009) **4**(12):e8233. doi:10.1371/journal.pone.0008233
16. Efroni S, Harel D, Cohen I. Emergent dynamics of thymocyte development and lineage determination. *PLoS Comput Biol* (2007) **3**(1):e13. doi:10.1371/journal.pcbi.0030013
17. Müller V, Bonhoeffer S. Quantitative constraints on the scope of negative selection. *Trends Immunol* (2003) **24**(3):132–5. doi:10.1016/S1471-4906(03)00028-0
18. Detours V, Mehr R, Perelson A. A quantitative theory of affinity-driven T cell repertoire selection. *J Theor Biol* (1999) **200**(4):389–403. doi:10.1006/jtbi.1999.1003
19. Prasad A, Zikherman J, Das J, Roose J, Weiss A, Chakraborty A. Origin of the sharp boundary that discriminates positive and negative selection of thymocytes. *Proc Natl Acad Sci U S A* (2009) **106**(2):528. doi:10.1073/pnas.0805981105
20. Ribeiro RM, Perelson AS. Determining thymic output quantitatively: using models to interpret experimental T-cell receptor excision circle (Trec) data. *Immunol Rev* (2007) **216**(1):21–34.
21. Sinclair C, Bains I, Yates AJ, Seddon B. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proc Natl Acad Sci U S A* (2013) **110**(31):E2905–14. doi:10.1073/pnas.1304859110
22. Surh CD, Sprent J. T-cell apoptosis detected in situ during positive and negative selection in the thymus. *Nature* (1994) **372**(6501):100–3. doi:10.1038/372100a0
23. Laufer TM, DeKoning J, Markowitz JS, Lo D, Glimcher LH. Unopposed positive selection and autoreactivity in mice expressing class II MHC only on thymic cortex. *Nature* (1996) **383**(6595):81–5. doi:10.1038/383081a0
24. van Meerwijk JP, Marguerat S, Lees RK, Germain RN, Fowlkes B, MacDonald HR. Quantitative impact of thymic clonal deletion on the T cell repertoire. *J Exp Med* (1997) **185**(3):377–84. doi:10.1084/jem.185.3.377
25. Merkenschlager M, Graf D, Lovatt M, Bommhardt U, Zamoyska R, Fisher AG. How many thymocytes audition for selection? *J Exp Med* (1997) **186**(7):1149–58. doi:10.1084/jem.186.7.1149
26. Stritesky GL, Xing Y, Erickson JR, Kalekar LA, Wang X, Mueller DL, et al. Murine thymic selection quantified using a unique method to capture deleted T cells. *Proc Natl Acad Sci U S A* (2013) **110**(12):4679–84. doi:10.1073/pnas.1217532110
27. Egerton M, Scollay R, Shortman K. Kinetics of mature T-cell development in the thymus. *Proc Natl Acad Sci U S A* (1990) **87**(7):2579–82. doi:10.1073/pnas.87.7.2579

28. Saini M, Sinclair C, Marshall D, Tolaini M, Sakaguchi S, Seddon B. Regulation of Zap70 expression during thymocyte development enables temporal separation of CD4 and CD8 repertoire selection at different signaling thresholds. *Sci Signal* (2010) **3**(114):ra23. doi:10.1126/scisignal.2000702
29. McCaughtry TM, Wilken MS, Hogquist KA. Thymic emigration revisited. *J Exp Med* (2007) **204**(11):2513–20. doi:10.1084/jem.20070601
30. Scollay RG, Butcher EC, Weissman IL. Thymus cell migration: quantitative aspects of cellular traffic from the thymus to the periphery in mice. *Eur J Immunol* (1980) **10**(3):210–8. doi:10.1002/eji.1830100310
31. Kenney JF, Keeping ES. *Mathematics of Statistics: Part One*. Princeton: van Nostrand (1954).
32. Kenney JF, Keeping ES. *Mathematics of Statistics: Part Two*. Princeton: van Nostrand (1962).
33. Mehr R, Perelson A, Fridkis-Hareli M, Globerson A. Feedback regulation of T cell development in the thymus. *J Theor Biol* (1996) **181**(2):157–67. doi:10.1006/jtbi.1996.0122
34. Scollay R, Godfrey D. Thymic emigration: conveyor belts or lucky dips? *Immunol Today* (1995) **16**(6):268–73. doi:10.1016/0167-5699(95)80179-0
35. Tough DF, Sprent J. Thymic emigration: conveyor belts or lucky dips? *Immunol Today* (1995) **16**(6):273–4. doi:10.1016/0167-5699(95)80180-4
36. Sim B-C, Aftahi N, Reilly C, Bogen B, Schwartz RH, Gascoigne NR, et al. Thymic skewing of the CD4/CD8 ratio maps with the T-cell receptor α-chain locus. *Curr Biol* (1998) **8**(12):701–S3. doi:10.1016/S0960-9822(98)70276-3
37. Boursalian TE, Golob J, Soper DM, Cooper CJ, Fink PJ. Continued maturation of thymic emigrants in the periphery. *Nat Immunol* (2004) **5**(4):418–25. doi:10.1038/ni1049
38. Spits H. Development of $\alpha\beta$ T cells in the human thymus. *Nat Rev Immunol* (2002) **2**(10):760–72. doi:10.1038/nri913
39. Marino JH, Tan C, Taylor AA, Bentley C, Van De Wiele CJ, Ranne R, et al. Differential IL-7 responses in developing human thymocytes. *Hum Immunol* (2010) **71**(4):329–33. doi:10.1016/j.humimm.2010.01.009
40. Molina-París C, Reynolds J, Lythe G, Coles MC. Mathematical model of naive t cell division and survival il-7 thresholds. *Front Immunol* (2013) **4**:434. doi:10.3389/fimmu.2013.00434
41. Allen LJS. *An Introduction to Stochastic Processes with Applications to Biology*. New Jersey: Pearson Education (2003).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 September 2013; accepted: 15 January 2014; published online: 14 February 2014.
*Citation: Sawicka M, Stritesky GL, Reynolds J, Abourashchi N, Lythe G, Molina-París C and Hogquist KA (2014) From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection. Front. Immunol. **5**:19. doi: 10.3389/fimmu.2014.00019*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Sawicka, Stritesky, Reynolds, Abourashchi, Lythe, Molina-París and Hogquist. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

A.1. STRINGENCY OF THYMIC SELECTION: A STOCHASTIC MODEL

In this section, we present the details that allow us to compute the stringency of thymic selection for the mathematical model considered in Section 2.2.

Let us assume that at time $t=0$, there exists a single T cell in a given compartment. In our case, the compartment can be the pre-DP, post-DP, or SP (either CD4 SP or CD8 SP) stages. The cell, at any time, may die (with rate, μ), divide (with rate, λ), and produce two daughter cells, or leave the compartment (with rate, ξ) to enter a different compartment. The waiting times for each event are assumed to be exponentially distributed, and daughter cells are assumed to behave identically to the initial single cell. We introduce the bivariate Markov process $\{X(t), Y(t)\}_{t\geq 0}$, where $X(t)$ is the number of cells in the compartment at time t, and $Y(t)$ is the number of cells which have left the compartment (to enter a different compartment). Our aim is to calculate the expected number of cells (and variance) that leave the compartment.

State probabilities for the Markov process are defined as follows (41)

$$p_{(x,y)}(t) = \text{Prob}\{X(t)=x, Y(t)=y | X(0)=1, Y(0)=0\}\,, \tag{A1}$$

and transition probabilities for this process are defined as follows (41)

$$\begin{aligned} p_{(w,z),(x,y)}(\Delta t) &= \text{Prob}\{X(t+\Delta t)=w, \\ &\quad Y(t+\Delta t)=z | X(t)=x, Y(t)=y\} \\ &= \begin{cases} \lambda x \Delta t + o(\Delta t), & \text{if } w=x+1, z=y\,, \\ \mu x \Delta t + o(\Delta t), & \text{if } w=x-1, z=y\,, \\ \xi x \Delta t + o(\Delta t), & \text{if } w=x-1, z=y+1\,, \\ 1-(\lambda+\mu+\xi)x\Delta t + o(\Delta t), & \text{if } w=x, z=y\,, \\ o(\Delta t), & \text{if } w,z \text{ otherwise}\,. \end{cases} \end{aligned} \tag{A2}$$

The Kolmogorov (or master) equation for this process is given by (41)

$$\begin{aligned} \frac{\mathrm{d}p_{(x,y)}(t)}{\mathrm{d}t} &= \lambda\,(x-1)p_{(x-1,y)}(t) + \mu\,(x+1)p_{(x+1,y)}(t) \\ &\quad + \xi\,(x+1)p_{(x+1,y-1)}(t) - (\lambda+\mu+\xi)\,x\,p_{(x,y)}(t)\,. \end{aligned} \tag{A3}$$

Let $m_X(t)$ be the expected number of cells in the compartment under consideration, and $m_Y(t)$ be the expected number of cells which have left the compartment. Similarly, let $m_{XX}(t)$ be the expectation of the random variable $X(t)^2$, $m_{XY}(t)$ be the expectation of the random variable $X(t)Y(t)$, and $m_{YY}(t)$ be the expectation of the random variable $Y(t)^2$. Then, making use of the probability generating function technique (41), we derive the time evolution for the first two moments of the system:

$$\frac{\mathrm{d}m_X(t)}{\mathrm{d}t} = (\lambda-\mu-\xi)m_X(t)\,, \tag{A4}$$

$$\frac{\mathrm{d}m_Y(t)}{\mathrm{d}t} = \xi\, m_X(t)\,, \tag{A5}$$

$$\frac{\mathrm{d}m_{XX}(t)}{\mathrm{d}t} = 2(\lambda-\mu-\xi)m_{XX}(t) + (\lambda-\mu-\xi)m_X(t)\,, \tag{A6}$$

$$\frac{\mathrm{d}m_{XY}(t)}{\mathrm{d}t} = (\lambda-\mu-\xi)m_{XY}(t) + \xi[m_{XX}(t) - m_X(t)]\,, \tag{A7}$$

$$\frac{\mathrm{d}m_{YY}(t)}{\mathrm{d}t} = \xi[2m_{XY}(t) + m_X(t)]. \tag{A8}$$

Given that we start with a single cell, the expected number of cells at time t is given by

$$m_X(t) = \mathrm{e}^{(\lambda-\mu-\xi)t}\,. \tag{A9}$$

Under the restriction $\lambda < \mu+\xi$, the expected number of cells tends to zero as $t\to+\infty$. This implies that all cells from the single T cell progenitor either die or leave the compartment for sufficiently large times. The expected number of cells which leave the compartment is given by

$$m_Y(t) = \frac{\xi}{\mu+\xi-\lambda}\left[1-\mathrm{e}^{(\lambda-\mu-\xi)t}\right]. \tag{A10}$$

As $t\to+\infty$, the expected number of cells to leave the compartment can be shown to be

$$\lim_{t\to+\infty} m_Y(t) = \frac{\xi}{\mu+\xi-\lambda}\,. \tag{A11}$$

We now solve the remaining ODEs equations (A6–A8), to find

$$\begin{aligned} m_{YY}(t) &= \frac{2\lambda\xi^2}{(\lambda-\mu-\xi)^3}\left[\frac{1}{2}\mathrm{e}^{2(\lambda-\mu-\xi)t} - \mathrm{e}^{(\lambda-\mu-\xi)t}\right] \\ &\quad - \frac{4\lambda\xi^2}{(\lambda-\mu-\xi)^2}\left(t - \frac{1}{\lambda-\mu-\xi}\right)\mathrm{e}^{(\lambda-\mu-\xi)t} \\ &\quad + \frac{\xi}{\lambda-\mu-\xi}\mathrm{e}^{(\lambda-\mu-\xi)t} - \frac{2\lambda\xi^2}{(\lambda-\mu-\xi)^3} \\ &\quad - \frac{\xi}{\lambda-\mu-\xi}\,. \end{aligned} \tag{A12}$$

It, therefore, follows that the random variable $Y(t)$, which represents the number of cells leaving the compartment under consideration, has the following variance (in the limit $t\to+\infty$)

$$\begin{aligned} \sigma_Y^2 &= \lim_{t\to\infty}\left[m_{YY}(t) - m_Y(t)^2\right] = \frac{2\lambda\xi^2}{(\mu+\xi-\lambda)^3} + \frac{\xi}{\mu+\xi-\lambda} \\ &\quad - \frac{\xi^2}{(\mu+\xi-\lambda)^2}\,. \end{aligned} \tag{A13}$$

A.2. STRINGENCY OF THYMIC SELECTION IN THE FOUR COMPARTMENT MODEL

The previous example can easily (but laboriously) be extended to the mathematical model introduced in Section 2.2.

We may evaluate the expected number of, for example, $CD4^+$ T cells produced by a single pre-DP progenitor (or more generally N pre-DP progenitors). We only present the time evolution of the moment generating function. The deterministic equations describing the mean and variance of the numbers of cells in each

compartment (pre-DP, post-DP, SP CD4, and SP CD8) are left for the reader to derive. When counting the number of $CD4^+$ T cells leaving the thymus, the moment generating function satisfies the following partial differential equation

$$\begin{aligned}\frac{\partial M}{\partial t} &= \mu_1(e^{-\theta_1}-1)\frac{\partial M}{\partial \theta_1} + \varphi_1(e^{-\theta_1}e^{\theta_2}-1)\frac{\partial M}{\partial \theta_1} \\ &+ (\mu_2+\varphi_8)(e^{-\theta_2}-1)\frac{\partial M}{\partial \theta_2} + \varphi_4(e^{-\theta_2}e^{\theta_4}-1)\frac{\partial M}{\partial \theta_2} \\ &+ \mu_4(e^{-\theta_4}-1)\frac{\partial M}{\partial \theta_4} + \lambda_4(e^{\theta_4}-1)\frac{\partial M}{\partial \theta_4} \\ &+ \xi_4(e^{-\theta_4}e^{\theta_8}-1)\frac{\partial M}{\partial \theta_4}\,. \end{aligned} \tag{A14}$$

The symmetry of the mathematical model implies that an equivalent equation for the number of $CD8^+$ T cells leaving the thymus can be obtained by interchanging the indexes 4 and 8. For our derived parameter set, the previous equation allows us to conclude that the expected number of $CD4^+$ T cells, a single thymocyte in the pre-DP compartment produces, is 0.069 (standard deviation 0.96), whereas the expected number of $CD8^+$ T cells which leave the thymus is 0.025 (standard deviation 0.59).

To put this into perspective, a population of 10^3 pre-DP thymocytes is expected to produce 69 $CD4^+$ T cells which leave the thymus (standard deviation 66), and 25 $CD8^+$ T cells (standard deviation 41). More generally, a population of N pre-DP thymocytes is expected to produce $0.069N$ $CD4^+$ T cells and $0.025N$ $CD8^+$ T cells.

frontiers in
IMMUNOLOGY

MINI REVIEW ARTICLE
published: 02 September 2013
doi: 10.3389/fimmu.2013.00264

Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis

***Gennady Bocharov*[1]*, *Tatyana Luzyanina*[2], *Jovana Cupovic*[3] *and Burkhard Ludewig*[3]**

[1] *Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow, Russia*
[2] *Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Pushchino, Russia*
[3] *Institute of Immunobiology, Kantonal Hospital St. Gallen, St. Gallen, Switzerland*

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Vitaly V. Ganusov, University of Tennessee, USA
Ollivier Hyrien, University of Rochester, USA

***Correspondence:**
Gennady Bocharov, Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina Street 8, Moscow 119333, Russia
e-mail: bocharov@inm.ras.ru

Flow cytometry-based analysis of lymphocyte division using carboxyfluorescein succinimidyl ester (CFSE) dye dilution permits acquisition of data describing cellular proliferation and differentiation. For example, CFSE histogram data enable quantitative insight into cellular turnover rates by applying mathematical models and parameter estimation techniques. Several mathematical models have been developed using different types of deterministic or stochastic approaches. However, analysis of CFSE proliferation assays is based on the premise that the label is halved in the two daughter cells. Importantly, asymmetry of protein distribution in lymphocyte division is a basic biological feature of cell division with the degree of the asymmetry depending on various factors. Here, we review the recent literature on asymmetric lymphocyte division and CFSE-based lymphocyte proliferation analysis. We suggest that division- and label-structured mathematical models describing CFSE-based cell proliferation should take into account asymmetry and time-lag in cell proliferation. Utilization of improved modeling algorithms will permit straightforward quantification of essential parameters describing the performance of activated lymphocytes.

Keywords: T cells, CFSE assay, asymmetric division, mathematical modeling

INTRODUCTION

The ability of the immune system to protect the host organism against live-threatening infections and tumors directly depends on the reactivity of lymphocytes to antigenic stimulation, with a key role of clonal T cell responses (1). The perception of infections as a race between the invading pathogen and immunity suggests that it is the knowledge of the proliferation and death rates of T cells which provides a quantitative basis for assessing the quality of the host immunity (2). For almost 20 years, flow cytometry-based analysis of intracellular fluorescent dye distribution has been used to assess the proliferative performance and differentiation patterns of lymphocytes (3–5). Since the prototype dye for this analysis is CFSE, the assay is commonly referred to as CFSE dilution assay or – more simply – CFSE assay. A quantitative characterization of T cell turnover which can be elaborated from time series of CFSE histograms ranges from "static" measures such as precursor cell frequency or mean generation number, to "dynamic" parameters characterizing the cell cycle progression and apoptosis rates (6). However, estimation of turnover parameters requires formulation of mathematical models of cell growth which can take various forms and differ in their complexity depending on the parameters of interest and the richness of the available data [comprehensively reviewed by De Boer and Perelson (7)]. Importantly, current approaches to the analysis of CFSE proliferation data are based on the assumption that cell division is symmetric, i.e., the fluorescent label is halved in the two daughter cells (3, 5, 7–9). However, a random and uneven partition of mass between the sister cells is considered as an axiom in cell biology since many years (10). Although the detailed knowledge of the intracellular reactions which affect the turnover and intracellular heterogeneity of CFSE labeled proteins is currently limited (11), it is broadly accepted that CFSE binds indiscriminately to intracellular proteins and the fluorescence intensity of any single cell is roughly proportional to the total number of CFSE molecules bound to proteins within that cell (12). The latter study proposed a method for the analysis of CFSE-labeling experiments which also considered the possibility of an unequal division of CFSE molecules between the daughter cells.

The inequality of the mass (protein) distribution to the daughter cells directly suggests that CFSE labeled proteins are unequally partitioned between daughter cells. Indeed, recent studies describing T cell activation showed that asymmetric cell division can be an inherent part of T cell growth and differentiation (13–16). However, direct experimental evidence for asymmetric partition of CFSE between daughter cells is still missing. Nevertheless, the existing deterministic mathematical frameworks should be amended to facilitate a quantitative analysis of CFSE-based lymphocyte proliferation when asymmetry of cell division associated with unequal partition of CFSE labeled proteins between the two daughter cell results in a poor resolution of divisional clusters in CFSE histograms. Here, we briefly summarize recent findings describing asymmetric lymphocyte division and progress in the analysis of CFSE-based lymphocyte activation. Moreover, cell proliferation is not an instantaneous process and it takes a finite time for a cell to progress from the G1-phase of the cell cycle to the completion of the M-phase. The duration of the continuous progression is called a time-lag and in general, needs to be explicitly parameterized in the model equations. Finally, we suggest that

mathematical models describing CFSE-based lymphocyte proliferation should consider both asymmetry in division and time-lag in proliferation.

ASYMMETRIC LYMPHOCYTE DIVISION

Symmetric or asymmetric cell divisions refer to the mode of cell division which results in two phenotypically identical- or different-daughter cells, respectively. The phenotypic features could be the cell size, cell surface receptors, intracellular components such as proteins (including those labeled with CFSE), transcription factors, or messenger RNA (17). Hence, these phenotypic differences provide the basis for the functional differences in the daughter cells, i.e., their cell fates.

Following encounter with their antigen displayed in the context of major histocompatibility complex molecules, naïve T lymphocytes go through well-orchestrated series of divisions generating different populations of cells that fulfill immediate effector functions or generate long-lived immunological memory. Two basic models explain the generation of such functionally distinct T cell phenotypes. According to the "one naïve cell – one fate" model, naïve lymphocytes are instructed to generate either effector or memory progeny (18). In this model, instruction of T cells, for example, is achieved through interaction with professional APCs (19). Hence, to preserve the instructing signal(s) received during activation and to maintain equality of the cells throughout division, T cells should divide in a symmetric fashion. The alternative model proposes asymmetric cell division as the mechanism that allows naïve T cells to give rise to two different daughter cells. These are referred to as proximal or distal daughter cell depending on their proximity to the APC. Such asymmetric T cell division represents a process that allows single cells to give rise to two, phenotypically and functionally different daughter cells, and thereby permits diversification of cell populations. In other words, one of the daughter cells inherits the potential to differentiate into full effector cell (proximal daughter), while the second daughter maintains the stemness of the mother cell. This principle feature of asymmetric cell division has also been described in developmental studies examining neurogenesis (20). Likewise, adaptation of adult tissues to changing environmental conditions such as the content of the gut requires rapid adaptation of one cell fraction while other cells maintain their high proliferative potential (21).

The processes involved in activation and differentiation of T cells, for example during infection have to swiftly generate cells with direct effector function to efficiently restrict viral replication (1). At the same time, some T cells should retain their ability to proliferate in order to prevent exhaustion of certain T cell subsets (22) and to facilitate generation of long-lived memory T cells (23). Indeed, Chang et al. (13) demonstrated that division of $CD8^+$ T cells specific for a viral peptide leads to the generation of daughter cells with different characteristics. CFSE-based assays revealed that asymmetry is established already during the first round of division and is dependent on the presence of the cognate antigen (13). Assessment of the protein content in the daughter cells generated during the first cell division showed that asymmetry established during mitosis is preserved throughout cytokinesis. Moreover, proximal and distal daughter cells exhibit different protein expression profiles and functional properties with proximal daughter cells exhibiting higher immediate protective capacity (13). The finding that proximal daughter cells exhibit higher CD8 co-receptor and LFA-1 expression facilitating formation of more frequent and longer lasting interactions with antigen presenting APCs (14) further emphasized that asymmetric division critically determines both T cell phenotype and function.

Asymmetric cell division is not only an important feature of $CD8^+$ T cell activation (13, 14), but also occurs during the activation and differentiation of $CD4^+$ T cells (13, 24) and B cells (25, 26). While naïve $CD8^+$ T cells require only one or only few encounters with APCs to proliferate and differentiate into effector cells, naïve $CD4^+$ T cells depend on multiple encounters in order to differentiate and to exhibit specialized effector functions (27). Hence, it is likely that $CD4^+$ T cells acquire their distinct phenotypes, e.g., Th1, Th2, or Th17, through multiple sequential asymmetric cell divisions. However, recent studies suggest that asymmetric cell division cannot be considered as the only mechanism that leads to the profound heterogeneity of T cell lineages (16). Thus, more research is required to resolve the contribution of sequential asymmetric T cell division to the generation of diverse T cell phenotypes. We suggest that a combination of CFSE-based T cell proliferation analysis with mathematical modeling may help – at least in part – to clarify this issue.

CURRENT MATHEMATICAL MODELS FOR CFSE-BASED LYMPHOCYTE PROLIFERATION ANALYSIS

Several mathematical models have been established for the analysis of CFSE-based proliferation assays (7, 9, 12, 28–32). The existing modeling frameworks can be subdivided on the basis of the major requirements for CFSE histogram data processing into two main categories (**Table 1**). The first group requires a decomposition of the CFSE histograms characterizing the distribution of cells with respect to the fluorescent dye into the distinct generations of cells. The procedure is based on fitting the CFSE histogram with a series of log-normal Gaussian distributions differing in their means and standard deviation and is implemented in commercially available standard software packages. Importantly, the assignment of distinct cell generations to CFSE clusters has remained an empiric process which depends heavily on initial labeling homogeneity, label degradation, cellular auto-fluorescence, and other factors including experimental skills of the researcher (33). As long as the division is symmetric (or almost symmetric) (**Figure 1A**), these factors can be tuned in a proper way to enable resolution of successive generations as distinct CFSE clusters (**Figure 1B**). Under these conditions a range of existing mathematical models can be tuned to estimate the turnover parameters of the stimulated lymphocyte population. The key features of the corresponding families of the models are outlined in **Table 1**, rows one to three. These models describe the population dynamics of cells which differ in the number of completed divisions and ignore the heterogeneity of the cells within a generation with respect to the CFSE content. The immunologically relevant issues that were addressed with the models of this group include regulatory effects of IL-2 on the T cell responses (34, 35), regulation of hematopoietic stem cells cycling (36), and kinetics of mouse erythroid progenitor cell differentiation (37).

Table 1 | Major features of mathematical models describing CFSE-based proliferation assays.

Cell proliferation model[1]	Input data	Estimated parameters	Mathematical approach[2]	Primary sources
A–B state model	Generation structure	Division entry-, apoptosis-rates, duration of division	DDE	Nordon et al. (6), Ganusov et al. (28)
G_0 model	Generation structure	Division entry-, apoptosis rates, duration of division	hPDE	Bernard et al. (51)
Random birth-death	Generation structure	Division-, apoptosis rates, progressor fraction	ODE, IE, branching processes	Ganusov et al. (28), Yates et al. (29), Lee et al. (30), Zilman et al. (31), Hyrien et al. (41), Veiga-Fernandes et al. (52), Revy et al. (53), Hawkins et al. (54)
Random birth-death, CFSE-structured	CFSE histograms	Division-, apoptosis-, CFSE decay rates	hPDE	Luzyanina et al. (38)
Random birth-death, generation-, CFSE-structured	CFSE histograms	Division-, apoptosis-, CFSE decay rates, auto-fluorescence	hPDE	Hasenauer et al. (40), Banks et al. (32)
Asymmetric division, G_0-model, generation-, CFSE-structured	CFSE histograms	Asymmetry, division-, apoptosis-, CFSE decay rates, time-lag of proliferation	hPDE	See text for details

[1]The following notations are used: "A–B state model" refers to the model of cell cycle in (55) in which the intermitotic period is composed of an A-state (major part of G1-phase) and a B-phase (conventional S, G2, and M phases); "G0 model" refers to the view of the cell cycle with two states (47), i.e., resting- (G0) and cycling-states (G1, S, G2, M). Conceptually, it is equivalent to the A–B state model. "Random birth-death" model refers to a discrete compartmental (generation structured) model of cycling cells (56). "Generation structured" refers to the mathematical model in which the cell population is decomposed into cohorts of cells which differ with respect to the number of completed division cycles; "CFSE-structured" represents the mathematical description of cell population in which the distribution (heterogeneity) of cells with respect to the fluorescence intensity is followed by considering the cell distribution function.

[2]DDE, delay differential equations; hPDE, hyperbolic partial differential equations; ODE, ordinary differential equations; IE, integral equations.

The second group of models which refer directly to the CFSE histograms seems to be more appropriate when the generational structure of the labeled population cannot be easily resolved (**Table 1**, rows four to five). The initially proposed model describes the evolution of the labeled cells distribution with respect to the CFSE level (38). Although this and similar models proved to be functional in estimating the proliferation- and death rates as functions of the structure variable directly from the histogram data (38, 39), the problem of translating the estimated functions into biologically meaningful parameters still requires the knowledge of the division structure of the lymphocyte population. A major breakthrough in the improvement of the distributed parameter models for the dynamics of heterogeneous CFSE labeled cell populations were recently proposed division- and label-structured mathematical models (32, 40). The major potential of this framework as an analytical tool is based upon the following features: (i) no need for CFSE histogram decomposition, (ii) characterization of cell growth in terms of generation dependent division- and death rates, (iii) an explicit form of the dependence of solution on the turnover parameters.

Another class of recently developed mathematical models which allow a direct fitting of the CFSE histograms is based on branching processes (12, 41). The approach allows for probabilistic characterization of cell activation, proliferation, and death from the CFSE dilution data and does not require the assumption about equality of CFSE division between the two daughter cells.

The first and the second group of models rely on the premise of symmetric cell division. However, tracing proliferation of other cell types such as cancer cells has been reported to be difficult (42) due to poorly resolved peaks of the different cell generations. Since cytokinesis is not perfect, it was suggested that the two daughter cells are unlikely to inherit exactly half of the CFSE fluorescence dye of the mother cell. An increase in the degree of the asymmetry of mass partition between daughter cells and hence disparate distribution of fluorescently labeled proteins should result in a poorer resolution of generational clusters as shown in **Figures 1C,D** for lower asymmetry and in **Figures 1E,F** for higher asymmetry. This in turn will lead to the generations overlap in CFSE histograms thus posing a limit to experimentalists' ability to resolve the individual generations using conventional decomposition methods.

MODELING ASYMMETRIC DIVISION OF CFSE LABELED T CELLS

We have been recently dealing with the analysis of the proliferative performance of monoclonal $CD8^+$ T cells recognizing an H2-Kb-binding epitope derived from the S protein of the mouse hepatitis virus (MHV). Clearance of MHV during acute infection is achieved through the combined action of type I interferons (43) and $CD8^+$ T cells (44). Moreover, $CD8^+$ T cells essentially contribute to control of the virus during persistent infection, for example in the central nervous system (45). We have initiated a project on the generation of avidity-tuned, antigen-specific T cells

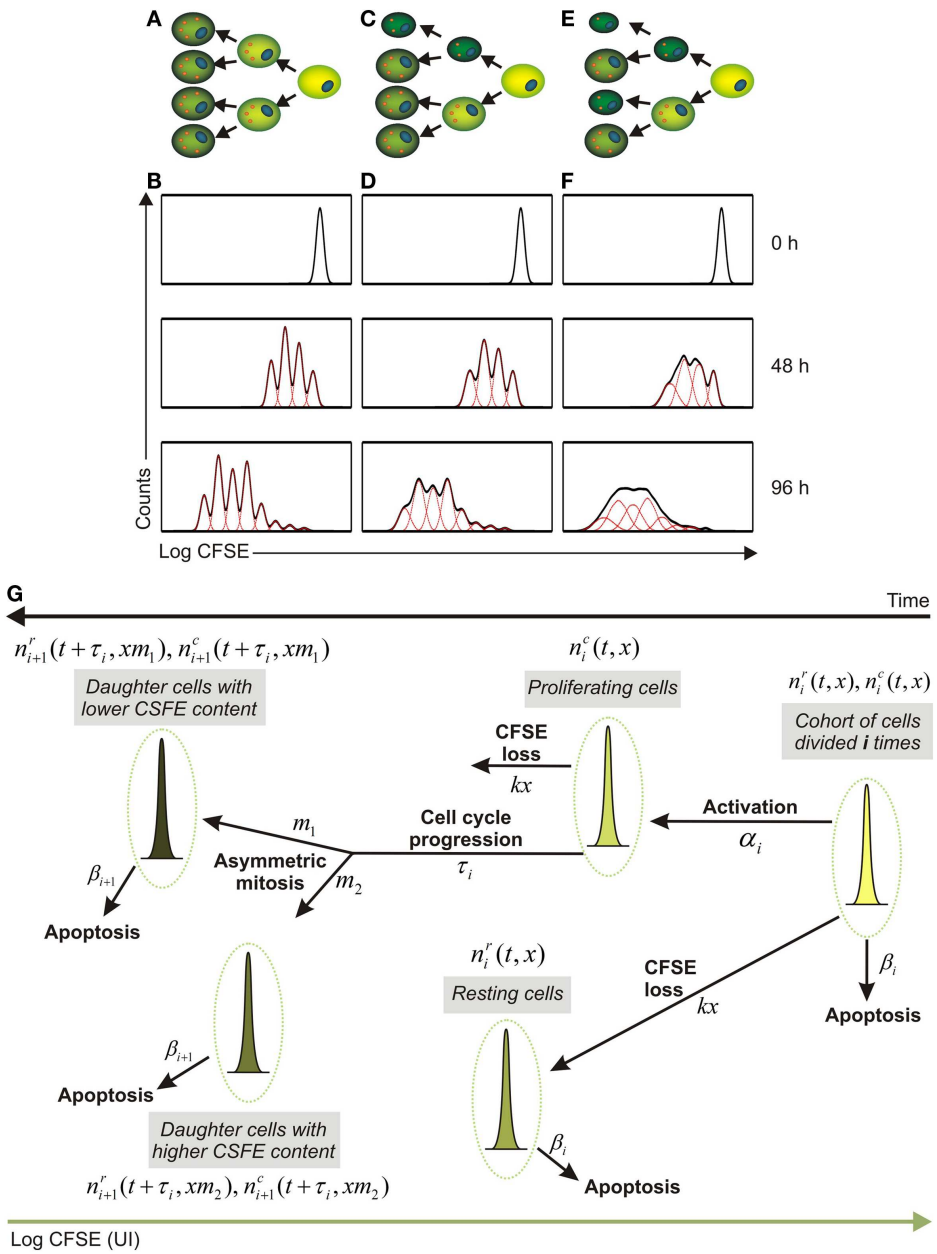


FIGURE 1 | Impact of asymmetry in T cell division impinges on fluorescent protein partition between daughter cells. **(A,B)** Symmetric cell division with equal distribution of the fluorescent dye between daughter cells **(A)** and modeled time course analysis of T cell proliferation as determined by flow cytometry [**(B)**, solid black lines]. Dashed red lines in **(B)** indicate the evolution of CFSE intensity of the cohorts (generations) of cell which differ in the number of completed divisions with the assumption of symmetric division. **(C,D)** Asymmetric cell division with "low" asymmetry **(C)** and modeled flow cytometric time course analysis of CFSE dilution [**(D)**, solid black lines] that corresponds to an asymmetry 46/54% [**(D)**, dashed red lines describe the CFSE distributions for cell cohorts differing in terms of the completed divisions]. **(E,F)** T cells dividing with "high" asymmetry **(E)** and corresponding model-generated flow cytometric CFSE dilution patterns [**(F)**, solid black lines] with asymmetry values of 42/58% describing the behavior of the T cells in this setting [**(F)**, dashed red lines describe the cell cohorts corresponding to different generations]. **(G)** Schematic representation of the structure of a mathematical approach which considers the division- and CFSE label-heterogeneity of proliferating cells as well as asymmetry and time of cell division. Some cells from the cohort of cells which completed "i" divisions are activated (α_i characterizes the activation rate) and progress through the cell cycle (τ_i stands for the duration of the progression through S-G_2-M phases), resulting to the generation of daughter cells which differ with respect to their CFSE content. Asymmetric mitosis refers to cell division which results into appearance of two phenotypically different daughter cells with a smaller and larger cell mass, respectively. These cells are characterized by an unequal amount of CFSE labeled proteins (m_1 and $m_2 = 1-m_1$, describe the fractions of CFSE from the mother cell inherited by the two daughter cells). The natural decay of the CFSE fluorescence intensity is taken into account (kx – stands for an exponential decay of CFSE loss).

for adoptive transfer as an option to augment antiviral immune responses during chronic infection. To this end, MHV-specific T cell receptors (TCRs) were cloned and tested in retrogenic systems (46). *In vitro* re-stimulation of the CFSE labeled monoclonal $CD8^+$ T cells showed that CFSE dilution was characterized by broadly varying patterns from highly distinct peaks to poorly resolved generational clusters. We propose that an explicit consideration of the asymmetry in protein partition between the daughter cells facilitates a consistent mathematical description of CFSE histogram time series data (**Figure 1G**). The appropriate mathematical framework should describe the population of CFSE labeled T cells by the distribution of cells with respect to CFSE amount (unit of intensity, UI). The subpopulations differ in terms of completed rounds of division and are further distinguished in resting and proliferating states, with the respective notation and i standing for the generation (number of completed divisions), t – for time and x – for CFSE amount per cell. A conceptual scheme of the modeling approach is shown in **Figure 1G** suggesting that such a model can be naturally formulated as an extension of a generation- and division-structured population balance model with the cell cycle represented according to the G_0 model (47) and the division asymmetry explicitly taken into account.

Under conditions of symmetric $CD8^+$ T cell division with the difference of protein partition between the sister cells being equal to zero (i.e., every daughter cell inherits half of the fluorescently labeled proteins of the mother cell), the model should predict clearly distinct generations (**Figure 1B**, dashed red lines). If the division is "weakly" asymmetric, i.e., the protein partition between the sister cells is different, the width of the CFSE distribution of the successive generations should become broader (**Figure 1D**, dashed red lines). Further increase in the degree of the asymmetry would result in a substantial overlap of the distinct cell generations (**Figure 1F**, dashed red lines). Obviously, this type of behavior of T cells – and other cells such as tumor cells needs to be regarded as a cause of a poor resolution of the generations in CFSE histograms (**Figures 1D,F**, solid black lines) thus creating an obstacle on the application of standard CFSE analysis tools.

The fitting of mathematical models for asymmetric cell division as conceptualized in **Figure 1G** to the time series data provides a tool for the estimation of the cell physiology parameters such as: (i) the generation-specific activation and death rates (α_i, β_i); (ii) the duration of the division cycle characterized by the time-lag (τ_i); (iii) the division asymmetry factors ($m_1 + m_2 = 1$), specifying the fraction of proteins which is inherited by the first and the second daughter cells, respectively; and (iv) the natural decay of the CFSE fluorescence intensity of the labeled cells (parameterized as kx). Taken together, asymmetric cell division improves assessment of T cell performance parameters from CFSE-based proliferation assays, even under conditions of poorly separated peaks.

CONCLUDING REMARKS

It is considered that the regulation of cell expansion and differentiation can occur by modulating the degree of asymmetry of cell divisions (17). It has been clearly shown that T lymphocyte division in response to pathogen exhibits unequal partitioning of proteins that mediate signaling and cell fate determination (13). Hence, asymmetric T lymphocyte division provides an additional mechanism for generating functionally heterogeneous populations of $CD8^+$ T cells both in primary and memory adaptive immune responses (48). Since a precise mechanistic link between the quantitative differences in partitioning of specific proteins between daughter cells and the developmental path of antigen-specific T cells remains to be established (49), mathematical modeling is now a key "instrument" for understanding the regulation of individual cell fates (15, 16, 50).

The addition of asymmetric T cell division to the analysis of CFSE-based proliferation data fills important gaps as it: (i) allows one to estimate the proliferation parameters for asymmetrically dividing cells directly from CFSE histograms with poorly resolved generations peaks and (ii) introduces a quantitative parameter which characterizes the difference in the partition of the fluorescent proteins between daughter cells and can be directly estimated from the same CFSE dilution data. A further question in CFSE analyses open for examination is the interplay between experimental variability, biological variability, and model parsimony. We expect that new mathematical tools for the analysis of a fundamental property of cell division, i.e., the phenotypic identity or differences of the daughter cells known as asymmetry, will be developed and introduced into daily experimental work. Thereby, a better understanding of the diversity and mechanisms underlying activation and homeostasis of T cell responses will be achieved.

ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation (130823 and 141918 to Burkhard Ludewig), the Vontobel Foundation (to Burkhard Ludewig), the Russian Foundation of Basic Research (11-01-00117a to Gennady Bocharov and Tatyana Luzyanina), and the Program of the Russian Academy of Sciences "Basic research for Medicine" (to Gennady Bocharov and Tatyana Luzyanina). We thank the reviewers and the editor for their insightful comments and the thorough work on our manuscript.

REFERENCES

1. Zinkernagel RM. Immunology taught by viruses. *Science* (1996) **271**:173–8. doi:10.1126/science.271.5246.173
2. Davenport MP, Belz GT, Ribeiro RM. The race between infection and immunity: how do pathogens set the pace? *Trends Immunol* (2009) **30**:61–6. doi:10.1016/j.it.2008.11.001
3. Lyons AB, Parish CR. Determination of lymphocyte division by flow cytometry. *J Immunol Methods* (1994) **171**:131–7. doi:10.1016/0022-1759(94)90236-4
4. Parish CR, Glidden MH, Quah BJ, Warren HS. Use of the intracellular fluorescent dye CFSE to monitor lymphocyte migration and proliferation. *Curr Protoc Immunol* (2009) **84**:4.9.1–4.9.13. doi:10.1002/0471142735.im0409s84
5. Quah BJ, Parish CR. New and improved methods for measuring lymphocyte proliferation in vitro and in vivo using CFSE-like fluorescent dyes. *J Immunol Methods* (2012) **379**:1–14. doi:10.1016/j.jim.2012.02.012
6. Nordon RE, Nakamura M, Ramirez C, Odell R. Analysis of growth kinetics by division tracking. *Immunol Cell Biol* (1999) **77**:523–9. doi:10.1046/j.1440-1711.1999.00869.x
7. De Boer RJ, Perelson AS. Quantifying T lymphocyte turnover. *J Theor Biol* (2013) **327**:45–87. doi:10.1016/j.jtbi.2012.12.025
8. Banks HT, Thompson WC. Mathematical models of dividing cell populations: application to CFSE data. *Math Model Nat Phenom* (2012) **7**:24–52. doi:10.1051/mmnp/20127504
9. Miao H, Jin X, Perelson AS, Wu H. Evaluation of multitype mathematical models for CFSE-labeling experiment data. *Bull Math Biol* (2012) **74**:300–26. doi:10.1007/s11538-011-9668-y

10. Sennerstam R. Partition of protein (mass) to sister cell pairs at mitosis: a re-evaluation. *J Cell Sci* (1988) **90**:301–6.
11. Banks HT, Choi A, Huffman T, Nardini J, Poag L, Thompson WC. Quantifying CFSE label decay in flow cytometry data. *Appl Math Lett* (2013) **26**:571–7. doi:10.1016/j.aml.2012.12.010
12. Hyrien O, Zand MS. A mixture model with dependent observations for the analysis of CFSE-labeling experiments. *J Am Stat Assoc* (2008) **103**:222–39. doi:10.1198/016214507000000194
13. Chang JT, Palanivel VR, Kinjyo I, Schambach F, Intlekofer AM, Banerjee A, et al. Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science* (2007) **315**:1687–91. doi:10.1126/science.1139393
14. King CG, Koehli S, Hausmann B, Schmaler M, Zehn D, Palmer E. T cell affinity regulates asymmetric division, effector cell differentiation, and tissue pathology. *Immunity* (2012) **37**:709–20. doi:10.1016/j.immuni.2012.06.021
15. Buchholz VR, Flossdorf M, Hensel I, Kretschmer L, Weissbrich B, Graf P, et al. Disparate individual fates compose robust CD8+ T cell immunity. *Science* (2013) **340**:630–5. doi:10.1126/science.1235454
16. Gerlach C, Rohr JC, Perie L, van RN, van Heijst JW, Velds A, et al. Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* (2013) **340**:635–9. doi:10.1126/science.1235487
17. Tajbakhsh S, Rocheteau P, Le R I. Asymmetric cell divisions and asymmetric cell fates. *Annu Rev Cell Dev Biol* (2009) **25**:671–99. doi:10.1146/annurev.cellbio.24.110707.175415
18. Ahmed R, Gray D. Immunological memory and protective immunity: understanding their relation. *Science* (1996) **272**:54–60. doi:10.1126/science.272.5258.54
19. Seder RA, Ahmed R. Similarities and differences in CD4+ and CD8+ effector and memory T cell generation. *Nat Immunol* (2003) **4**:835–42. doi:10.1038/ni969
20. Chenn A, McConnell SK. Cleavage orientation and the asymmetric inheritance of Notch1 immunoreactivity in mammalian neurogenesis. *Cell* (1995) **82**:631–41. doi:10.1016/0092-8674(95)90035-7
21. Edgar BA. Intestinal stem cells: no longer immortal but ever so clever. *EMBO J* (2012) **31**:2441–3. doi:10.1038/emboj.2012.133
22. Probst HC, Tschannen K, Gallimore A, Martinic M, Basler M, Dumrese T, et al. Immunodominance of an antiviral cytotoxic T cell response is shaped by the kinetics of viral protein expression. *J Immunol* (2003) **171**:5415–22.
23. Wherry EJ, Ahmed R. Memory CD8 T-cell differentiation during viral infection. *J Virol* (2004) **78**:5535–45. doi:10.1128/JVI.78.11.5535-5545.2004
24. Choi YS, Kageyama R, Eto D, Escobar TC, Johnston RJ, Monticelli L, et al. ICOS receptor instructs T follicular helper cell versus effector cell differentiation via induction of the transcriptional repressor Bcl6. *Immunity* (2011) **34**:932–46. doi:10.1016/j.immuni.2011.03.023
25. Barnett BE, Ciocca ML, Goenka R, Barnett LG, Wu J, Laufer TM, et al. Asymmetric B cell division in the germinal center reaction. *Science* (2012) **335**:342–4. doi:10.1126/science.1213495
26. Duffy KR, Wellard CJ, Markham JF, Zhou JH, Holmberg R, Hawkins ED, et al. Activation-induced B cell fates are selected by intracellular stochastic competition. *Science* (2012) **335**:338–41. doi:10.1126/science.1213230
27. Celli S, Garcia Z, Bousso P. CD4 T cells integrate signals delivered during successive DC encounters in vivo. *J Exp Med* (2005) **202**:1271–8. doi:10.1084/jem.20051018
28. Ganusov VV, Pilyugin SS, De Boer RJ, Murali-Krishna K, Ahmed R, Antia R. Quantifying cell turnover using CFSE data. *J Immunol Methods* (2005) **298**:183–200. doi:10.1016/j.jim.2005.01.011
29. Yates A, Chan C, Strid J, Moon S, Callard R, George AJ, et al. Reconstruction of cell population dynamics using CFSE. *BMC Bioinformatics* (2007) **8**(196):196. doi:10.1186/1471-2105-8-196
30. Lee HY, Hawkins E, Zand MS, Mosmann T, Wu H, Hodgkin PD, et al. Interpreting CFSE obtained division histories of B cells in vitro with Smith-Martin and cyton type models. *Bull Math Biol* (2009) **71**:1649–70. doi:10.1007/s11538-009-9418-6
31. Zilman A, Ganusov VV, Perelson AS. Stochastic models of lymphocyte proliferation and death. *PLoS ONE* (2010) **5**:e12775. doi:10.1371/journal.pone.0012775
32. Banks HT, Thompson WC, Peligero C, Giest S, Argilaguet J, Meyerhans A. A division-dependent compartmental model for computing cell numbers in CFSE-based lymphocyte proliferation assays. *Math Biosci Eng* (2012) **9**:699–736. doi:10.3934/mbe.2012.9.699
33. Ko KH, Odell R, Nordon RE. Analysis of cell differentiation by division tracking cytometry. *Cytometry A* (2007) **71**:773–82.
34. Deenick EK, Gett AV, Hodgkin PD. Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. *J Immunol* (2003) **170**:4963–72.
35. Ganusov VV, Milutinovic D, De Boer RJ. IL-2 regulates expansion of CD4+ T cell populations by affecting cell death: insights from modeling CFSE data. *J Immunol* (2007) **179**:950–7.
36. Takizawa H, Regoes RR, Boddupalli CS, Bonhoeffer S, Manz MG. Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. *J Exp Med* (2011) **208**:273–84. doi:10.1084/jem.20101643
37. Akbarian V, Wang W, Audet J. Measurement of generation-dependent proliferation rates and death rates during mouse erythroid progenitor cell differentiation. *Cytometry A* (2012) **81**:382–9. doi:10.1002/cyto.a.22031
38. Luzyanina T, Roose D, Schenkel T, Sester M, Ehl S, Meyerhans A, et al. Numerical modelling of label-structured cell population growth using CFSE distribution data. *Theor Biol Med Model* (2007) **4**(26):26. doi:10.1186/1742-4682-4-26
39. Banks HT, Sutton KL, Thompson WC, Bocharov G, Roose D, Schenkel T, et al. Estimation of cell proliferation dynamics using CFSE data. *Bull Math Biol* (2011) **73**:116–50. doi:10.1007/s11538-010-9524-5
40. Hasenauer J, Schittler D, Allgower F. Analysis and simulation of division- and label-structured population models: a new tool to analyze proliferation assays. *Bull Math Biol* (2012) **74**:2692–732. doi:10.1007/s11538-012-9774-5
41. Hyrien O, Chen R, Zand MS. An age-dependent branching process model for the analysis of CFSE-labeling experiments. *Biol Direct* (2010) **5**(41):41–5. doi:10.1186/1745-6150-5-41
42. Matera G, Lupi M, Ubezio P. Heterogeneous cell response to topotecan in a CFSE-based proliferation test. *Cytometry A* (2004) **62**:118–28. doi:10.1002/cyto.a.20097
43. Cervantes-Barragan L, Kalinke U, Zust R, Konig M, Reizis B, Lopez-Macias C, et al. Type I IFN-mediated protection of macrophages and dendritic cells secures control of murine coronavirus infection. *J Immunol* (2009) **182**:1099–106.
44. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol* (2009) **7**:439–50. doi:10.1038/nrmicro2147
45. Bergmann CC, Lane TE, Stohlman SA. Coronavirus infection of the central nervous system: host-virus stand-off. *Nat Rev Microbiol* (2006) **4**:121–32. doi:10.1038/nrmicro1343
46. Uckert W, Schumacher TN. TCR transgenes and transgene cassettes for TCR gene therapy: status in 2008. *Cancer Immunol Immunother* (2009) **58**:809–22. doi:10.1007/s00262-008-0649-4
47. Burns FJ, Tannock IF. On the existence of a G0-phase in the cell cycle. *Cell Tissue Kinet* (1970) **3**:321–34.
48. Ciocca ML, Barnett BE, Burkhardt JK, Chang JT, Reiner SL. Cutting edge: asymmetric memory T cell division in response to rechallenge. *J Immunol* (2012) **188**:4145–8. doi:10.4049/jimmunol.1200176
49. Oliaro J, Van Ham V, Sacirbegovic F, Pasam A, Bomzon Z, Pham K, et al. Asymmetric cell division of T cells upon antigen presentation uses multiple conserved mechanisms. *J Immunol* (2010) **185**:367–75. doi:10.4049/jimmunol.0903627
50. Buchholz VR, Graf P, Busch DH. The smallest unit: effector and memory CD8(+) T cell differentiation on the single cell level. *Front Immunol* (2013) **4**(31):31. doi:10.3389/fimmu.2013.00031
51. Bernard S, Pujo-Menjouet L, Mackey MC. Analysis of cell kinetics using a cell division marker: mathematical modeling of experimental data. *Biophys J* (2003) **84**:3414–24. doi:10.1016/S0006-3495(03)70063-0
52. Veiga-Fernandes H, Walter U, Bourgeois C, McLean A, Rocha B. Response of naive and memory CD8+ T cells to antigen stimulation in vivo. *Nat Immunol* (2000) **1**:47–53. doi:10.1038/76907
53. Revy P, Sospedra M, Barbour B, Trautmann A. Functional antigen-independent synapses formed between T cells and dendritic cells. *Nat Immunol* (2001) **2**:925–31. doi:10.1038/ni713

54. Hawkins ED, Turner ML, Dowling MR, van GC, Hodgkin PD. A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc Natl Acad Sci U S A.* (2007) **104**:5032–7. doi:10.1073/pnas.0700026104
55. Smith JA, Martin L. Do cells cycle? *Proc Natl Acad Sci U S A* (1973) **70**:1263–7. doi:10.1073/pnas.70.4.1263
56. Kendall DG. On the role of variable generation time in the development of a stochastic birth process. *Biometrika* (1948) **35**:316–30. doi: 10.2307/2332354

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 June 2013; accepted: 19 August 2013; published online: 02 September 2013.
*Citation: Bocharov G, Luzyanina T, Cupovic J and Ludewig B (2013) Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis. Front. Immunol. **4**:264. doi: 10.3389/fimmu.2013.00264*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Bocharov, Luzyanina, Cupovic and Ludewig. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

PERSPECTIVE ARTICLE
published: 01 October 2013
doi: 10.3389/fimmu.2013.00300

Dynamical and mechanistic reconstructive approaches of T lymphocyte dynamics: using visual modeling languages to bridge the gap between immunologists, theoreticians, and programmers

Véronique Thomas-Vaslin[1,2]*, Adrien Six[1,2], Jean-Gabriel Ganascia[3,4] and Hugues Bersini[5]

[1] *UPMC Univ Paris 06, UMR7211, Immunology, Immunopathology, Immunotherapy (I3),Integrative Immunology, Paris, France*
[2] *CNRS, UMR 7211, Immunology-Immunopathology-Immunotherapy (I3) Integrative Immunology, Paris, France*
[3] *UPMC Univ Paris 06, UMR 7606, ACASA-LIP6, Paris, France*
[4] *CNRS, UMR 7606, ACASA-LIP6, CNRS, Paris, France*
[5] *Université Libre de Bruxelles, IRIDIA-Code, Bruxelles, Belgium*

Edited by:
Rob J. de Boer, Utrecht University, Netherlands

Reviewed by:
Christopher E. Rudd, University of Cambridge, UK
Yoram Louzoun, Bar-Ilan University, Israel
Hillel Kugler, Microsoft Research, UK

****Correspondence:***
Véronique Thomas-Vaslin, UMR 7211, Immunology, Immunopathology, Immunotherapy, Integrative Immunology, UPMC-CNRS, 83 Boulevard de l'Hôpital, Paris 75013, France
e-mail: veronique.thomas-vaslin@upmc.fr

Dynamic modeling of lymphocyte behavior has primarily been based on populations based differential equations or on cellular agents moving in space and interacting each other. The final steps of this modeling effort are expressed in a code written in a programing language. On account of the complete lack of standardization of the different steps to proceed, we have to deplore poor communication and sharing between experimentalists, theoreticians and programmers. The adoption of diagrammatic visual computer language should however greatly help the immunologists to better communicate, to more easily identify the models similarities and facilitate the reuse and extension of existing software models. Since immunologists often conceptualize the dynamical evolution of immune systems in terms of "state-transitions" of biological objects, we promote the use of unified modeling language (UML) state-transition diagram. To demonstrate the feasibility of this approach, we present a UML refactoring of two published models on thymocyte differentiation. Originally built with different modeling strategies, a mathematical ordinary differential equation-based model and a cellular automata model, the two models are now in the same visual formalism and can be compared.

Keywords: state-transition diagram, computer modeling, cell dynamics, agent-based model, complex system

The perspective is to encourage immunologists involved into mathematical modeling or software productions, to adopt a visual graphical language, here mainly the unified modeling language (UML) "state-transition" diagram to ease the communication, the reuse and the extension of their models.

COMPLEXITY OF THE IMMUNE SYSTEM

The immune system is a complex biological adaptive, highly diversified, robust and resilient system, characterized by complexity at different levels. Lymphocytes are the central actors of the immune system, in the middle of a multi-scale biological organization, "from molecule to organism". Multi-scale modeling remains a challenge, as for other biological systems (1). Despite recent systems biology initiatives to understand and model the immune system (2), we are still far from having the appropriate tools to understand its dynamics and to easily communicate among various researchers who observe this system at different levels of granularity and attempt through software modeling to answer different questions. Several complementary experimental methods and models have been used to explore lymphocyte dynamics and turnover (3, 4) and to model it in health, aging and diseases (5).

DRAWBACKS OF CURRENT DYNAMICS LYMPHOCYTE MODELING AND EVOLUTION

System dynamics models deal with time, formalized with two distinct concepts as "discrete time," by a succession of time points and intervals, or as "continuous time" (6). Models of lymphocyte population dynamics and turnover (3) have primarily been based on mechanistic reconstruction with continuous time models. The fluxes of cell populations are then described by differential equations. These mathematical models describe for example the thymocyte differentiation and selection (7), until the thymic export (8–10), the homeostasis of CD4/CD8 T cells (11, 12), the CD8 immune response (13, 14), or the Bromodeoxyuridine or deuterium labeling (15) to account for turnover. Up to now, simulations and validation of some of these models reveal interesting T cell dynamics properties: how the system grows, self-maintains as well as the effects of perturbations, i.e. how the system reacts to antigens, collapses and reorganizes. However, integrating the heterogeneity of cell populations, phenotypes, lineages, cell location and interactions, cell differentiation across generations (16) in the different biological, and time scales, is problematic in such a mathematical form, which make these models particularly difficult to handle.

The evolution of homogeneous mathematical model of cell populations (17) toward "spatialized," discrete, and heterogeneous software models (18) has allowed the reproduction and observation of more detailed and thus complex behaviors. For example, this made possible to model lymphocyte dynamics from thymic selection (19, 20) up to quantitative modeling of immune responses, as extensively reviewed (21) with development of agent-based and automata models (22). However, both population-based mathematical model (a top-down approach) and discrete cell-based model (a bottom-up approach) and the various platforms developed have limitations (23). Conversely, the Statecharts language (24) and the visual reactive tools (25) such as biocharts (26) and reactive animation applied to various systems (27) developed by Harel et al. are a powerful way to simulate complex dynamical biological behavior with more didactic representation than equations. Such models have revealed emergent properties during thymic differentiation (28) pancreatic islet organogenesis but also the immune response in lymph node (29).

LACK OF INTEROPERABILITY, UNDER-USE OF SOFTWARE MODELS

Although in Immunology there is more than 20-years tradition of software and mathematical modeling, very few of them have been the object of further exploitation once published and made available (30). Models are often under-used because experimentalists can be reluctant to entertain mathematical formalization and because published models are largely disposable: rapidly forgotten after being published, instead of providing a foundation to build upon. Moreover, the various expressions of these models with different mathematical descriptions, programing languages, software libraries and graphical packages, require much effort in understanding and running the software and prevent interoperability.

USING VISUAL LANGUAGE TO COMMUNICATE AND EXECUTE MODELS

Immunologists often conceptualize the dynamical evolution of their systems in terms of "state-transitions" of biological objects and do it by means of personalized and informal graphical illustrations. Thus, the adoption of a more formal and standard type of state-transition diagram could improve the current situation to not only help biologists to better understand each other but also to facilitate the production and the reading of software code executing these visual transitions, at level of populations or agents (31).

Thus, in this paper, we promote the development and usage of a visual, computational approach more comprehensible than mathematical equations and programing instructions. This should improve our understanding of lymphocyte dynamics, the exchange on this understanding and simplify the implementation of models by non-specialists delivered from the production of executable code or mathematical equations, to concentrate to *in silico* experiments.

LEVEL OF ABSTRACTION AND MULTI-SCALE MODELING

A model describes a complex system from the "real word" and thus requires abstraction. This abstraction is performed as the immunologist decides about an experimental protocol in order to observe selected objects at different scales and to follow them in time and space. For example, the capacities of the immune system to preserve the homeostasis and to provide rapid adaptation to an antigen and anamnestic responses can be observed at the organism level, through physiological or pathological clinical observations that relate to lower scale levels. At molecular level, the somatic generation of the diversity of an immuno-receptor, as the TCR, allows for a dynamic network of interactions with antigens. At the cell level, this leads to clonal selection, activation and division. At the organ level, the fluidity of the system insures constant tissue redistribution of cells and molecules, cell migration from thymus to spleen and lymph nodes.

Thus, models of lymphocyte population dynamics and turnover consist in reconstructing the components or "entities" of the system across various scales, from molecules to organisms, to determine the relations/interactions through space (varying from micrometer to meters) and "processes" through time (varying between microseconds to years) as explained below. However, the formalization and abstraction of the immune objects as entities undergoing processes, with the help of spatial and dynamic ontologies, respectively defined as SNAP and SPAN (32), as well as cell/molecule interactions (33), is rarely done, maintaining a language-barrier between biologists and theoreticians. In the following, some examples will be given to help the immunologists with the transition between current mathematical models to computer ones and with the terms currently used in modeling.

DEFINE ENTITIES, STATES, LOCATION, INTERACTIONS, GRANULARITY

The immune "entities" could be described according to the language used by the modeler. A cell exists in one "state": it could be quiescent or in a given phase of the cell cycle or dead. In addition the phenotype and/or a function of a cell define a given state, as CD4 helper T cells. Cells are "located" in various tissues and are in "relation" with other entities. Finally, cells can be considered at various level of "granularity." For example, T lymphocyte populations are "aggregation" of T cells according to criteria of phenotype, structure or function, although heterogeneity still prevails inside these populations at lower granularity. Accordingly, cells can be modeled at population level (with continuous model as ordinary equation) or at cell level according to space (with discrete model as automata or multi-agent system).

DEFINE PROCESSES

According to ontologies, cells participate to various processes, such as division, activation, differentiation, interaction, clonotype selection, apoptosis or migration. According to the states of the cells, their evolution can thus be modeled as "state-transition" that can be applied to various processes in parallel: for example, a thymocyte can differentiate while migrating in cortex and medulla. Note that processes at other levels like molecular or organ levels can similarly be described and modeled. Finally, all these process will determine the global cell dynamics and turnover.

THE UNIFIED MODELING LANGUAGE FOR HIGH-LEVEL MODELING

"High-level programing languages" are based on abstraction and use of natural language that is easier to understand as compared to "low level programing language," based on codes. Thus, visual modeling language considers biological-object as conceptual abstract-objects that endure processes. The level of abstraction

allowed by these diagrams makes possible to distinguish more easily the "entities" as T cells and the "processes" that occur at different levels such as differentiation, migration and cell cycle. Moreover, such "state-transition diagrams" allow computing parallel pathways at various scales to avoid redundancy that is inherent in the formal description of multi-level, heterogeneous and concurrent systems and to model heterogeneity in a very simplified and economical form (as compared to mathematical equations) (31).We have thus used the well-established Unified Modelling Language (UML – a sofware standard) that still remains approachable to the lab-immunologist, convenient for the theoretician and that can be directly adopted for the high-level graphical depiction (31, 34). The adoption of UML state-transition diagrams that transcends any programing language or computer platform, will allow both experimentalists and theorists to work together at a higher level than writing software code or mathematical equations. This final step is progressively more and more automatized out of the diagrams. Example of basic transformations of mathematical equations into state-transition diagrams including elementary, parallel, independent, or coupled state transition have been given (31), to familiarize biologists with the general approach.

REFACTORING FROM LOW LEVEL (CODE, EQUATIONS) TO HIGH-LEVEL (DIAGRAM) MODELING LANGUAGES

To convince the immunologist of the feasibility of this approach as well as the benefit gained by adopting it, we sketch in the rest of the paper how existing "low level programed" immune models should gain in readability and accessibility by adopting a "high-level graphical" representation under the form of "state-transition diagrams." We present a "refactoring" of two published models of T cell biology in the thymus. Refactoring consists in restructuring the code or equations of a model to improve its expression, readability and extensibility, without changing its external behavior. One model consists in cell population differentiation modeling with differential equations (continuous model). The other one is a discrete model. Originally it was an automata model consisting of a discrete lattice, where each site (cell) in a given state, follows some rules in space and time that depends on local neighbors (18). It has been refactorized as an agent-based model (ABM), depicting individual cell behavior through thymus differentiation and migration. It would be much too long and redundant to describe in details the behavior of these two models. We do not pretend here to modify at all the results obtained by the running of these models (the readers interested in these results are invited to access the original papers). We have just reshape them into a state-transition diagrammatic form that allows execution of simulations reproducing the original results with similar parameter values.

POPULATION-BASED MODEL DESCRIBING THE CONVEYOR-BELT T CELL DIFFERENTIATION IN THYMUS

The original model (8) is a compartmentalized ordinary differential equation (ODE) model, rather complex to read and manage by immunologists. This model reflects the conceptual "conveyor belt" model of thymic T cell differentiation, schematically represented by immunologists by the continuous ordered transition of cells through the different stages with time (35–37). **Figure 1** represents a biological schema, originally published and the "state-transition" description of the model (in a UML state-transition diagram) as it is proposed now. Although the original model is composed of 30 differential equations, the whole mathematical description and the code that captures it, can easily be deduced and regenerated from the **Figure 1**. Conversely, the mathematical equations can be automatically generated from the state-transition diagram as previously described (38). In essence, the model is summarized by the input, transition and output from the thymus, by "parallel processes" that occur concomitantly, as differentiation, cell cycle, proliferation and death, and by exit from the thymus. Note that these parallel processes concern various biological levels and time scales. The "differentiation" process represents each stage of the conveyor belt, from double negative (DN) to single positive (SP) cells, with flows into and from a particular stage according to the general equation:

$$dx_i/dt = 2\gamma\ px_{i-1} - \big(p + d + u\,(i)\big)x_i$$

p, d, and u represent proliferation, death, and differentiation, respectively, and x_i represents the ith stage. The model mainly consists of constant hematopoietic progenitor influx in thymus (Sn); differentiation between thymocyte developmental phenotypes DN and double-positive (DP) cells differentiation into $CD4^+$ or $CD8^+$ cells, then egression of SP stage, either, to the periphery [Us4(i), Us8(i)]; proliferation [Pn(y), Pp(y), and Ps(y)]; positive and negative selection (a4, a8); and natural cell death (Dn, Dp, and Ds). In parallel, the "cell cycle" is represented: the cell switches between quiescence (G0) and cycle with division (S/M). The parameter γ set to 1 represents the cell division into two daughter cells. There is the possibility to induce a perturbation into the system through the specific depletion of T cells entering the S/M cycle phase (p), if γ is set to 0. This represents the presence or absence of a pharmacogenetic conditional treatment by ganciclovir that induces apoptosis related to the incorporation of a nucleotide analog during DNA elongation. This rule applies to all cell populations except in late DP quiescent cells as indicated in the schema. The "proliferation" depicts that the daughters of a proliferating cell transit into the next generational compartment, except during treatment ($\gamma = 0$) when dividing cells die by apoptosis and are lost from the model. The parameter u is an increasing function of generation (G1 to n), making cells more likely to differentiate between phenotypic compartments as they progress through the cell generations.

The parallelism in this graphical model largely simplifies the original formulation of ODEs while remaining faithful to it. Hierarchy and compound states are present again clearly reducing the diagram clutter. Other representation of the model depicting the differentiation with linear cell generations is also possible (31).

DISCRETE MODEL DESCRIBING THE DIFFERENTIATION ALONG THE MIGRATION OF T CELLS IN THE THYMUS IN A 2-D ENVIRONMENT

The original model (20) is a discrete-based "cellular automata" computer model. The model depicts the behavior of individual thymocytes that evolve in the 2-D epithelial cell network, guided by chemokines gradients. The current model (**Figure 2**) is now an Agent-Based Model (ABM). Again, the interested reader is referred to the original paper for a detailed understanding of the simulation. Although available for download, the 40 pages of FORTRAN source code are far from easy to understand. After refactoring, the transition rules of any agent (thymocytes) map onto a parallel

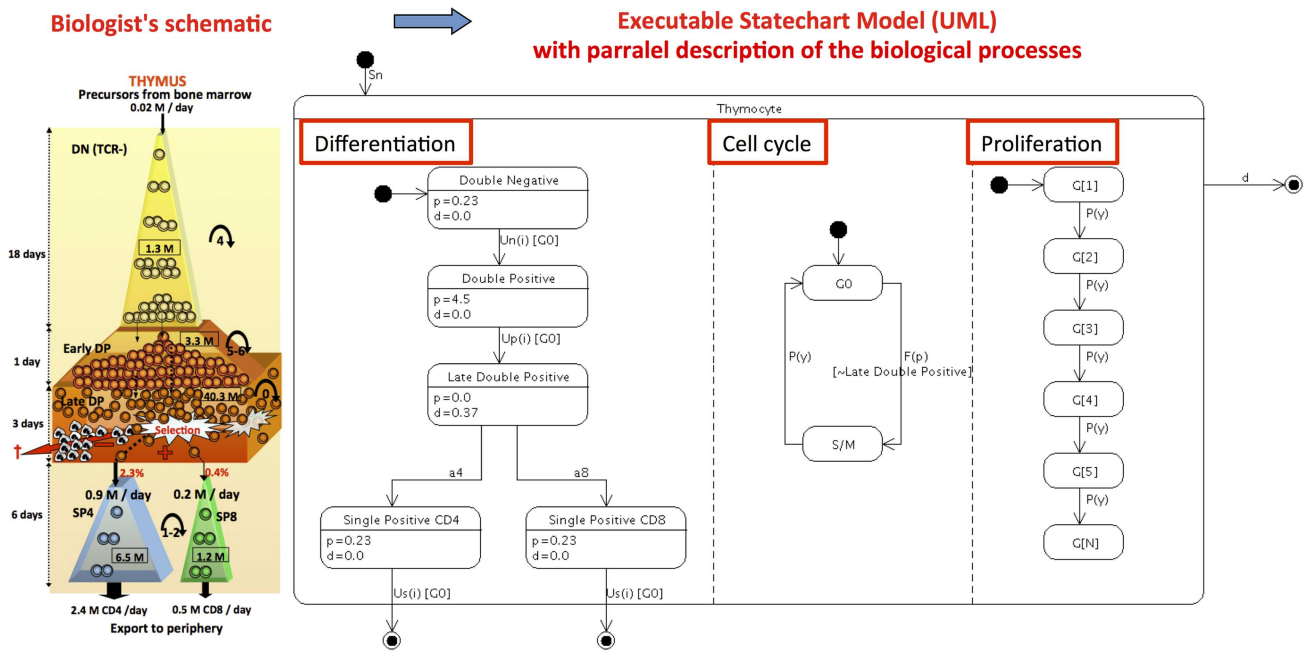


FIGURE 1 | Refactoring the continuous population (ODE) conveyor-belt model of thymocyte differentiation (8) into a computer executable visual language: The «biological schema »of thymocyte dynamics originally proposed is transformed into a UML state-transition diagram. This diagram represents the evolution of cell populations in the thymus, represented by their state from DN to SP, with UML figuration of the input (close circle arrow) and output representing death and thymus exit (open circle arrow) and transitions (oriented arrows). Parallel processes (underlined with red box) as differentiation, cell cycle and proliferation are depicted more explicitly in this representation than with the original 30 mathematical equations. Annotation with the proliferation and death rates values indicated in each state are values from the best scenario observed in the original paper.

state-transition diagram. The conception of the state-transition diagram, as done here, should considerably improve the understanding of the model (even for the original programmer) allowing the researchers to progress further with the existing simulator. A complete description of the model includes additional implementation details that are abstractions of the mechanisms behind cell decisions to differentiate. The parallel state-transition diagram represents the different simultaneous transitions taking place in the model and coded as various cellular automata rules: a cell in the model as it differentiates transits in successive states, it may be bound or not to thymic epithelial cells via TCR/MHC, it moves and may be located into one of several anatomical compartments of the thymus. As indicated in the boxes, the gradient of chemokines (k) orients the migration of cells for each specific stage of differentiation, and chemokines are localized in specific areas of the thymus. A T cell sums up the time and number of interactions with the same or different epithelial cells. This sum value determines whether the T cell is positively or negatively selected. DP and SP phenotypes have their own threshold parameters. They cannot differentiate until they cross a threshold number of interactions. If, after a given time, a DP cell has not reached this threshold, it enters apoptosis by neglect. If the time is too long, it is negatively selected. A threshold parameter simulates the phenotype decision. With this "signal-duration" hypothesis, long duration TCR-MHC interactions promote the CD4 phenotype and short duration promotes the CD8 phenotype.

It is important to notice that both refactored models, population-based and ABM can now be compared, are directly executable and can provide simulations of physiology, pathologies, and treatment, while not being the scope of this paper. Moreover, their parameters can be automatically tuned to fit experimental data. Any ABM model can also be run as a population version to save time in simulation (McEwan et al., manuscript in preparation). Moreover, the flexibility of these diagrams allows assembling the parts of the biological puzzle piece after piece and improving the models.

PERSPECTIVES

As shown here, state-transition diagrams can represent high-level semantics suitable to clarify immunological concepts and to aid communication among interdisciplinary researchers. It can also represent low levels quantitative information suitable for individual-based ABM and population-based ODE modeling. Organization of immune knowledge using a standardized, diagrammatic formal language should greatly improve knowledge integration at multi-scale levels and sharing between experimentalist and theoretician collaborators, rendering their software more readable, scalable, and usable. We are currently working on ways of automatically generating executable code out of these state-transition diagrams. State-transition diagrams supports the extension and interoperability of published models. This will help for dynamic computational modeling of lymphocyte behavior

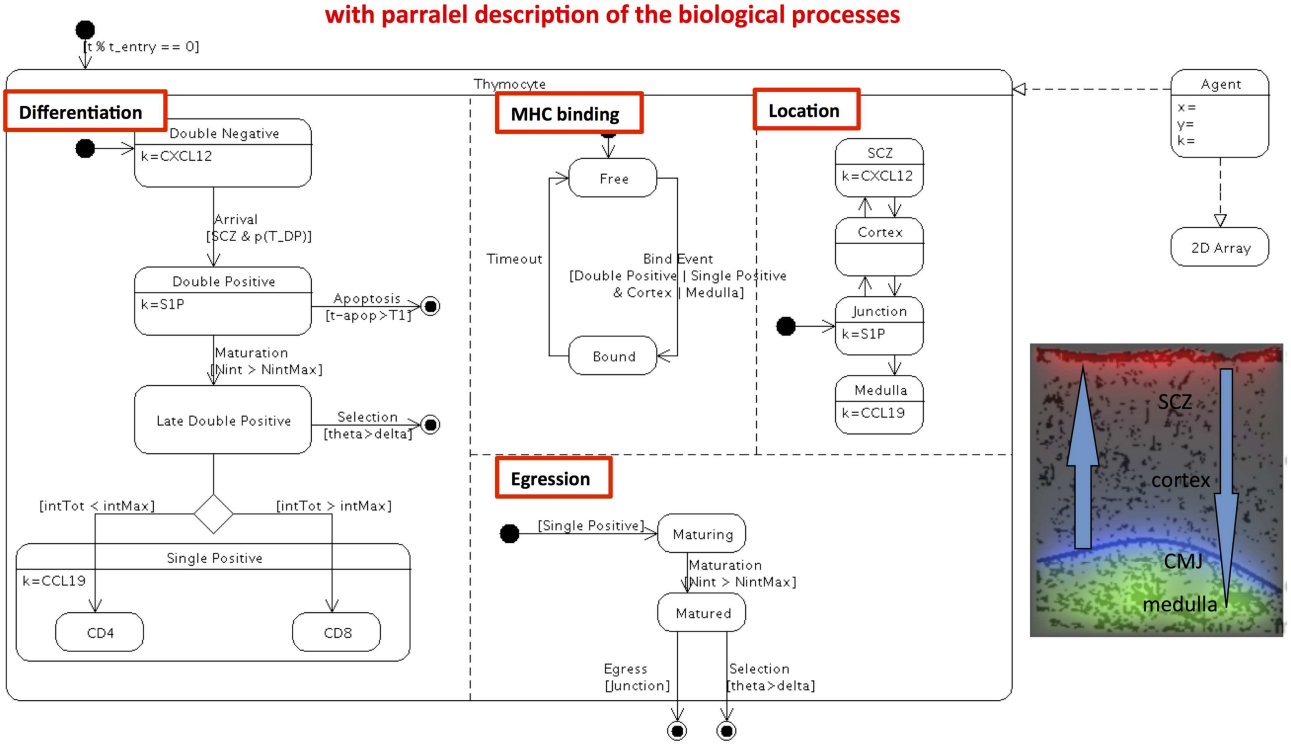


FIGURE 2 | Refactoring the automata single-cell model of thymocyte differentiation and migration (20) into a computer executable visual language. The 40 pages of Fortran code is transformed in computer UML state-transition diagram. This diagrams describes the experimentally observable heterogeneity, and the biologically relevant parallel processes (underlined in red box). As in **Figure 1**, input, output, and oriented transitions are described in the state-transition diagram. Parallel state-transitions represent the evolution of single cells in the thymus, represented by their differentiation from DN to SP stage, sequential binding event of TCR/MHC peptide on epithelial cells, thymic location, egression of cell when matured. Additional qualitative abstractions for computational model of individual cells in semi-realistic environment are represented by the 2-D array: cells are allowed migrating sequentially through the epithelial cell network (black network) across the various thymic areas, guided by chemokine gradients CXCL12 (red), CCL19/CCL21 (green), S1P (blue).

in health and diseases, and for "*in silico*" experiments to predict and explain the puzzling T cells dynamics and the effect of immunological perturbations.

AUTHOR CONTRIBUTIONS

Véronique Thomas-Vaslin and Hugues Bersini designed the project and wrote the article, Adrien Six and Jean-Gabriel Ganascia participated to discussions.

ACKNOWLEDGMENTS

The authors thank C. McEwan for computer modeling developments and discussions, B. Zerath, B. Horsmans for early refactoring of models, W. Chaara, and E. Mariotti-Ferrandiz from critical reading of the manuscript. Hugues Bersini is an invited Professor at UPMC. Financial supports: Emergence-UPMC program, PHC Tournesol EGIDE, CNRS PEPS BMI, Investissements d'Avenir program (ANR-11-IDEX-0004-02), LabEx Transimmunom.

REFERENCES

1. Lavelle C, Berry H, Beslon G, Ginelli F, Giavitto J, Kapoula Z, et al. From molecules to organisms: towards multiscale integrated models of biological systems. *Theor Biol Insights* (2008) **1**:13–22.
2. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser ID. Systems biology in immunology: a computational modeling perspective (*). *Annu Rev Immunol* (2011) **29**:527–85. doi:10.1146/annurev-immunol-030409-101317
3. Thomas-Vaslin V, Six A, Bellier B, Klatzmann D. Lymphocytes dynamics repertoires, modeling. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. *Encyclopedia of Systems Biology*. New York: Springer (2013). p. 1149–62. doi:10.1007/978-1-4419-9863-7_96
4. Thomas-Vaslin V, Six A, Bellier B, Klatzmann D. Lymphocyte dynamics and repertoire, biological methods. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. *Encyclopedia of Systems Biology*. New York: Springer (2013). p. 1145–49. doi:10.1007/978-1-4419-9863-7_95
5. Asquith B, Borghans JA, Ganusov VV, Macallan DC. Lymphocyte kinetics in health and disease. *Trends Immunol* (2009) **30**(4):182–9. doi:10.1016/j.it.2009.07.013
6. Ossimitz G, Mrotzek M. The basics of system dynamics: discrete vs. continuous modelling of time. *Proceedings of the 26th International Conference of the System Dynamics Society*. Wiley-Blackwell (July 2008) (2008). Available from: http://teoriasistemas.mex.tl/imagesnew/6/6/4/5/2/SimulaciondiscretavsContinua.pdf
7. Mehr R, Globerson A, Perelson AS. Modeling positive and negative selection and differentiation processes in the thymus. *J Theor Biol* (1995) **175**(1):103–26. doi:10.1006/jtbi.1995.0124
8. Thomas-Vaslin V, Altes HK, de Boer RJ, Klatzmann D. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *J Immunol* (2008) **180**(4):2240–50.

9. Bains I, Thiebaut R, Yates AJ, Callard R. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**(7):4329–36. doi:10.4049/jimmunol.0900743
10. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, de Boer AB, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* (2012) **36**(2):288–97. doi:10.1016/j.immuni.2012.02.006
11. Mehr R, Perelson AS. Blind T-cell homeostasis and the CD4/CD8 ratio in the thymus and peripheral blood. *J Acquir Immune Defic Syndr Hum Retrovirol* (1997) **14**(5):387–98. doi:10.1097/00042560-199704150-00001
12. Almeida AR, Amado IF, Reynolds J, Berges J, Lythe G, Molina-París C, et al. Quorum-sensing in CD4+ T cell homeostasis: a hypothesis and a model. *Front Immunol* (2012) **3**:125. doi:10.3389/fimmu.2012.00125
13. Chao DL, Davenport MP, Forrest S, Perelson AS. A stochastic model of cytotoxic T cell responses. *J Theor Biol* (2004) **228**(2):227–40. doi:10.1016/j.jtbi.2003.12.011
14. Terry E, Marvel J, Arpin C, Gandrillon O, Crauste F. Mathematical model of the primary CD8 T cell immune response: stability analysis of a nonlinear age-structured system. *J Math Biol* (2012) **65**(2):263–91. doi:10.1007/s00285-011-0459-8
15. De Boer RJ, Perelson AS, Ribeiro RM. Modelling deuterium labelling of lymphocytes with temporal and/or kinetic heterogeneity. *J R Soc Interface* (2012) **9**(74):2191–200. doi:10.1098/rsif.2012.0149
16. Yamanaka YJ, Gierahn TM, Love JC. The dynamic lives of T cells: new approaches and themes. *Trends Immunol* (2013) **34**(2):59–66. doi:10.1016/j.it.2012.10.006
17. Louzoun Y. The evolution of mathematical immunology. *Immunol Rev* (2007) **216**:9–20.
18. Celada F, Seiden PE. A computer model of cellular interactions in the immune system. *Immunol Today* (1992) **13**:56. doi:10.1016/0167-5699(92)90135-T
19. Morpurgo D, Serentha R, Seiden PE, Celada F. Modelling thymic functions in a cellular automaton. *Int Immunol* (1995) **7**(4):505–16. doi:10.1093/intimm/7.4.505
20. Souza-e-Silva H, Savino W, Feijoo RA, Vasconcelos AT. A cellular automata-based mathematical model for thymocyte development. *PLoS One* (2009) **4**(12):e8233. doi:10.1371/journal.pone.0008233
21. Cohn M, Mata J. Quantitative modeling of immune responses. *Immunol Rev* (2007) **216**(1):5–8.
22. Chavali AK, Gianchandani EP, Tung KS, Lawrence MB, Peirce SM, Papin JA. Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. *Trends Immunol* (2008) **29**(12):589–99. doi:10.1016/j.it.2008.08.006
23. Bianca C, Pennisi M. Immune system modelling by top-down and bottom-up approaches. *Int Math Forum* (2012) **7**(3):109–28.
24. Harel D. Statecharts: a visual formalism for complex systems. *Sci Comput Program* (1987) **8**(3):231–74. doi:10.1016/0167-6423(87)90035-9
25. Efroni S, Harel D, Cohen IR. Toward rigorous comprehension of biological complexity: modeling, execution, and visualization of thymic T-cell maturation. *Genome Res* (2003) **13**(11):2485–97. doi:10.1101/gr.1215303
26. Kugler H, Larjo A, Harel D. Biocharts: a visual formalism for complex biological systems. *J R Soc Interface* (2010) **7**(48):1015–24. doi:10.1098/rsif.2009.0457
27. Vainas O, Harel D, Cohen IR, Efroni S. Reactive animation: from piecemeal experimentation to reactive biological systems. *Autoimmunity* (2011) **44**(4):271–81. doi:10.3109/08916934.2010.523260
28. Efroni S, Harel D, Cohen IR. Emergent dynamics of thymocyte development and lineage determination. *PLoS Comput Biol* (2007) **3**(1):e13. doi:10.1371/journal.pcbi.0030013
29. Swerdlin NI, Cohen IR, Harel D. The lymph node B cell immune response: dynamic analysis in-silico. *Proc IEEE* (2008) **96**(8):1421–43. doi:10.1109/JPROC.2008.925435
30. Bersini H, editor. Object-oriented refactoring of existing immune models. In: *Artificial Immune Systems. Lecture notes in Computer Science*. Berlin: Springer-Verlag (2009). p. 27–40.
31. Bersini H, Klatzmann D, Six A, Thomas-Vaslin V. State-transition diagrams for biologists. *PLoS One* (2012) **7**(7):e41165. doi:10.1371/journal.pone.0041165
32. Grenon P, Smith B. SNAP and SPAN: towards dynamic spatial ontology. *Spat Cogn Comput* (2004) **4**(1):69–104. doi:10.1207/s15427633scc0401_5
33. Pappalardo F, Lefranc MP, Lollini PL, Motta S. A novel paradigm for cell and molecule interaction ontology: from the CMM model to IMGT-ONTOLOGY. *Immunome Res* (2010) **6**(1):1. doi:10.1186/1745-7580-6-1
34. Bersini H. UML for ABM. *J Artif Soc Soc Simul* (2012) **15**(1):
35. Penit C, Vasseur F. Sequential events in thymocyte differentiation and thymus regeneration revealed by a combination of bromodeoxyuridine DNA labeling and antimitotic drug treatment. *J Immunol* (1988) **140**(10):3315–23.
36. Penit C, Lucas B, Vasseur F. Cell expansion and growth arrest phases during the transition from precursor (CD4-8-) to immature (CD4+8+) thymocytes in normal and genetically modified mice. *J Immunol* (1995) **154**(10):5103–13.
37. Scollay R, Godfrey D. Thymic emigration: conveyor belts or lucky dips? *Immunol Today* (1995) **16**:268–74. doi:10.1016/0167-5699(95)80179-0
38. McEwan CH, Bersini H, Klatzmann D, Thomas-Vaslin V, Six A. A computational technique to scale mathematical models towards complex heterogeneous systems. *COSMOS Workshop ECAL 2011 Conference*; Paris: Luniver Press (2011).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 June 2013; paper pending published: 07 July 2013; accepted: 09 September 2013; published online: 01 October 2013.
*Citation: Thomas-Vaslin V, Six A, Ganascia J-G and Bersini H (2013) Dynamical and mechanistic reconstructive approaches of T lymphocyte dynamics: using visual modeling languages to bridge the gap between immunologists, theoreticians, and programmers. Front. Immunol. **4**:300. doi: 10.3389/fimmu.2013.00300*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Thomas-Vaslin, Six, Ganascia and Bersini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 November 2013
doi: 10.3389/fimmu.2013.00366

A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children

Tharindi Hapuarachchi, Joanna Lewis and Robin E. Callard*

Institute of Child Health and CoMPLEX, University College London, London, UK

Edited by:
Miles Davenport, University of New South Wales, Australia

Reviewed by:
Grant Lythe, University of Leeds, UK
Mark Dowling, Walter and Eliza Hall Institute, Australia

****Correspondence:***
Robin E. Callard, Institute of Child Health and CoMPLEX, University College London, 30 Guilford Street, London WC1N 1EH, UK
e-mail: r.callard@ucl.ac.uk

The size and composition of the T lymphocyte compartment is subject to strict homeostatic regulation and is remarkably stable throughout life in spite of variable dynamics in cell production and death during T cell development and immune responses. Homeostasis is achieved by careful orchestration of lymphocyte survival and cell division. New T cells are generated from the thymus and the number of peripheral T cells is regulated by controlling survival and proliferation. How these processes combine is however very complex. Thymic output increases in the first year of life and then decreases but is crucial for establishing repertoire diversity. Proliferation of new naive T cells plays a crucial role for maintaining numbers but at a potential cost to TCR repertoire diversity. A mechanistic two-compartment model of T cell homeostasis is described here that includes specific terms for thymic output, cell proliferation, and cell death of both resting and dividing cells. The model successfully predicts the homeostatic set point for T cells in adults and identifies variables that determine the total number of T cells. It also accurately predicts T cell numbers in children in early life despite rapid changes in thymic output and growth over this period.

Keywords: naive T cells, homeostasis, CD4 T cells, mathematical modeling, mechanistic modeling, children

INTRODUCTION

The naive T cell compartment in humans is generated early in development by the thymus and then maintained throughout life by continued export from the thymus and cell division in the periphery. In adult humans, the naive T cell compartment is comprised of roughly 10^{11} cells circulating between the blood and the peripheral lymphoid organs. It is estimated to comprise at least 10^8 different T cell receptor specificities (1) providing a broad spectrum of protection in a diverse pathogen environment. The size and composition (T cell receptor diversity) of the naive T cell compartment are subject to strict homeostatic regulation and are remarkably stable throughout adult life despite changing rates of cell production and death during T cell development and immune responses (2, 3). Homeostasis is achieved by control of lymphocyte survival and cell division. Naive T cell survival and peripheral cell division depends on access to the cytokine IL7 (4–7) and TCR signals (8, 9) through contact with self-peptide MHC (spMHC) on dendritic cells (10). In lymphoreplete mice, naive T cells are largely non-cycling (11) whereas homeostatic cell division plays an important role in maintaining naive T cell homeostasis in humans, where cell division is evident in the naive pool (12, 13).

In children, homeostatic control of the T cell compartment may be affected by both the growth of the child with the accompanying increased blood volume (14) and changes in thymic output, which increases to a maximum over the first year of life and then declines to reach an approximately steady level by the age of about 20 years (15). As a result, the CD4 naive T cell count (cells/μl) in children declines over the first 10–20 years of life whereas the total number of naive CD4 cells increases as the child grows (**Figure 1**). This raises important questions about whether the homeostatic mechanisms themselves change during early life or whether the numbers of naive CD4 T cells observed are determined only by the changes in thymic output and growth.

To date, our understanding of the processes controlling survival and proliferation of T cells has been largely qualitative and detailed quantitative knowledge of how homeostatic responses result in the observed equilibrium of the T cell pool with a given size and composition is lacking. Here, a two-compartment mathematical model of homeostasis is presented incorporating specific terms for thymic export into the naive CD4 compartment, rates of entry into cell division and death (survival) rates for both the resting and dividing cell compartments. In this sense, the model can be considered as mechanistic in comparison to empirical or descriptive models where the parameters have no direct biological meaning. The results illustrate the importance of T cell dynamics for the maintenance of constant naive CD4 T cell numbers in adults and the growth of the T cell compartment in children.

MATERIALS AND METHODS

A MODEL OF NAIVE T CELL HOMEOSTASIS

T cell homeostasis can be described using a two-compartment model of resting and dividing cells with input from the thymus into the resting compartment as shown in **Figure 2** (16, 17). In this model we will consider only naive CD4 T cells assuming no antigenic stimulation and maturation of naive to memory cells. The same model could in principle also be applied to memory cells and CD8 T cells.

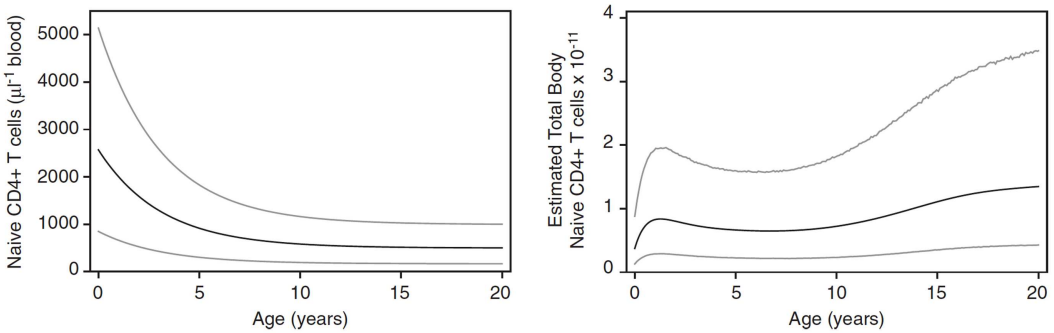


FIGURE 1 | Changes in naive CD4 T cell concentration and total whole body numbers with age. Taken from Bains et al. (14).

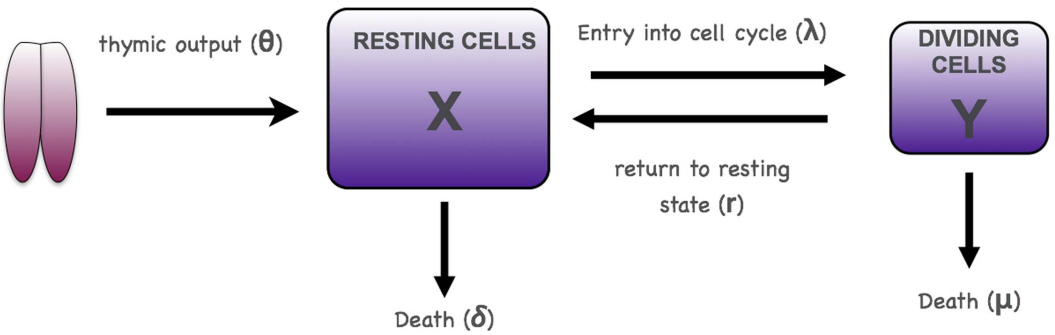


FIGURE 2 | Scheme for two-compartment model of homeostasis.

This model can be expressed mathematically using two coupled ordinary non-linear differential equations:

$$\frac{dX}{dt} = \theta + 2rY - \lambda_{(X+Y)}X - \delta_{(X+Y)}X$$
$$\frac{dY}{dt} = \lambda_{(X+Y)}X - rY - \mu_{(Y)}$$

where X is the number of non-dividing (resting) T cells and Y the number of cells undergoing cell division (**Figure 2**). The parameter θ represents T cell export from the thymus, λ the rate at which resting cells enter cell division, r the rate at which dividing cells return to the resting state, δ the death rate of resting cells, and μ the death rate of dividing cells. λ, r, δ, and μ are all first order rate constants, in units of day^{-1}, whereas θ is a zero-order constant, in units of cells day^{-1}.

To develop this model, it is important to have biologically appropriate forms for each of these parameters. Thymic output is known to vary with age with a maximum at about 1 year, which then declines rapidly until about 20 years of age and more slowly thereafter (15, 18). The value of θ for a 20 year old has been estimated to be 3×10^8 CD4 T cells day^{-1} (15). This value was used for modeling CD4 T cell homeostasis in a young adult. In children, the value of θ changes rapidly with age. An expression for θ from 0 to 20 years was determined as described previously (15).

An appropriate form for the rate of entry into cell cycle λ is (19)

$$\lambda = \lambda_0 \exp\left[-N(t)/\varepsilon\right]$$

where $N(t) = X(t) + Y(t)$, i.e., the total number of T cells at time t.

This expression is based on competition between resting naive T cells for signals to enter cell division: TCR signaling by self-peptide MHC and resources such as IL7 (4, 5, 8, 9, 20, 21). The term λ_0 represents the intrinsic ability of a T cell to respond under conditions of no competition (very few cells or an unlimited supply of homeostatic proliferative signals such as IL7), ε is proportional to the amount of resource (IL7) available and N is the total number of T cells competing for the resource. The rate of entry into cell cycle therefore deceases exponentially with decreasing resource or increasing cell number. The rate at which dividing cells return to the resting state r is determined by the length of time taken for one division [known to be about 6 h (19)] and experimental evidence that in homeostatic cell division cells return to the resting state after one division (19, 22). The death rate μ of activated T cells takes the form $\mu = \mu' Y$, which represents density-dependent AICD (activated induced cell death) by Fas–Fas ligand interactions (23, 24). Finally, the death rate of resting cells δ takes the form

$$\delta = \delta_0 \exp\left[N(t)/\rho\right].$$

Table 1 | Parameter values used for the model.

Parameter	Description	Value
Θ	Thymic output for adult 20 years old	3×10^8 cells day^{-1} (15)
λ_0	Rate of entry into cell cycle with infinite resource	0.055 cell^{-1} day^{-1}
ε	Resource for entry into cell cycle	1
δ_0	Death rate of resting cells with infinite resource	0.02 cell^{-1} day^{-1}
ρ	Resource for resting cell survival	100
r	Rate of return from dividing to resting state	4 day^{-1} (every 6 h)
μ	Death rate of dividing cells	15 day^{-1}

Similar to λ, this term is also derived from the reported dependence of cell survival on competition for a survival signal such as IL7 (resource) where δ_0 is the intrinsic ability of a cell to die under conditions of no competition (very few cells or an unlimited supply of the survival signal) and ρ is proportional to the amount of available resource providing the survival signal (IL7) (21). Parameter values used in the model are shown in **Table 1**.

The model was solved numerically using NDSolve, the proprietary numerical ODE solver in Mathematica that automatically selects the most appropriate method and adapts the step size so that the estimated errors are within the specified tolerance.

RESULTS

HOMEOSTATIC SET POINT IN ADULTS

The homeostatic set point for adults was examined by testing the behavior of the model starting with cell numbers well below and above the equilibrium and with an adult thymic output of 3×10^8 cells day^{-1} (15). Initial conditions were 0 dividing cells and either 0.01 or 2 ($\times 10^{11}$) resting cells. As shown in **Figure 3**, a stable equilibrium of total naïve CD4 T cell numbers (resting plus dividing) was obtained at just over 10^{11} cells, after 200–300 days (see also **Figure 1**). A Jacobian analysis showed that the solutions were stable over a wide range of parameter values for r (>0.281), μ' (<106.79), ε (<1.01), and λ_0 ($>1.05 \times 10^{-14}$) and stability did not depend on thymic output (θ), δ_0, or ρ.

The ratio of dividing to resting cells is shown in **Figure 4**. With lymphopenic starting conditions of 0.01 ($\times 10^{11}$) resting cells, the proportion of proliferating cells (blue curve) increased very rapidly from 0 to 0.013 and then slowly declined over about 200 days to reach an equilibrium at about 0.5%. In contrast, under starting conditions of excessive T cells, the ratio of dividing to non-dividing cells increased rapidly at first from 0 to about 0.2% and then slowly to reach the same equilibrium of about 0.5%. This equilibrium point is consistent with a low level of cell division in the naive compartment of adult humans as reported previously (25, 26).

EFFECTS OF COMPETITION FOR SURVIVAL AND DIVISION SIGNALS

Next, we investigated the effect of the amount of resource available for cell division (ε) and cell survival (ρ) (**Figure 5**). Consistent with competition between naive CD4 T cells for a resource such as spMHC and/or IL7 in order to survive and undergo cell division, the number of cells at homeostatic equilibrium decreased as the resource terms ε for proliferation and ρ for survival decreased. Interestingly, the rate of entry into cell cycle was significantly more sensitive than survival to changes in resource concentration, consistent with different thresholds for proliferation and survival as previously described (27).

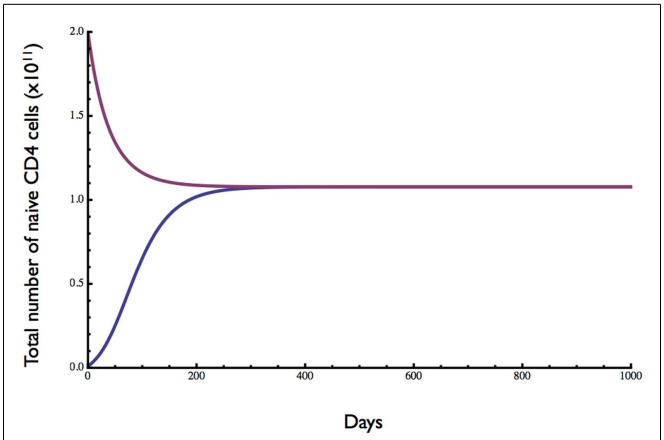


FIGURE 3 | Dynamics of naive CD4 T cell homeostasis in adults predicted by the model. Starting with 0.01- or 2-fold the approximate number of naive CD4 T cells in a replete young adult, an equilibrium of about 10^{11} cells is reached within 200–300 days. Parameter values are given in **Table 1**.

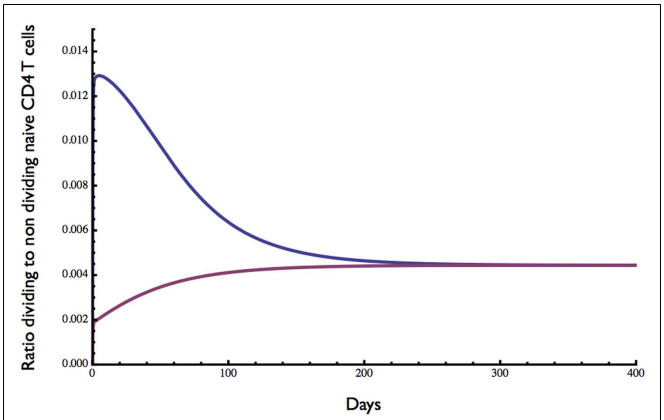


FIGURE 4 | Ratio of dividing to resting cells predicted by the model. Starting with 0.01×10^{11} (blue curve) or 2×10^{11} (red curve) the approximate ratio of dividing to resting cells changes over time to reach an equilibrium of about 0.4%. As expected, the proportion of proliferating cells is greater when the initial cell number is low. Parameter values used in the model are the same as in **Figure 3**.

T CELL HOMEOSTASIS IN CHILDREN

Having established the behavior of the two-compartment model for naive CD4 T cell homeostasis in adults, we sought to determine whether it could also be used to explain the changes in T cell numbers that occur during childhood. During the first few years

of life as thymic output changes and children grow, the naive CD4 T cell concentration decreases while the total number increases (**Figure 1**). The question is whether the homeostatic mechanism described by the model is in itself enough to explain these variations, or whether different and changing mechanisms apply in children. To examine this question, the model was used to predict the concentration of naive CD4 T cells in cells/μl of blood by converting total numbers to concentration using the estimated blood volume of children at different ages (14). In addition, the changes in thymic output that occur over the first few years of life with a peak at 1 year and then a decline (15) were incorporated into the model. The prediction from the model was then simply compared without parameter fitting to data collected from a cohort of healthy children (born to HIV infected mothers) from the European Collaborative Study on HIV infected pregnant women and their children (28) (**Figure 6**). As can be seen, the model predicted the concentration of T cells over the first 3 years of life extremely well suggesting that the homeostatic mechanisms in children and adults are essentially the same with the only difference being thymic output and growth with a concomitant increase in blood volume.

DISCUSSION

The two-compartment mathematical model presented here is based on the known biology of naive T cell homeostasis. It is derived from an earlier simple model that ignored thymic output and competition for resources (17, 24). Although only naive CD4 T cells are considered here, the same model would essentially be applicable to CD8 T cells. Naive single positive CD4 T cells enter the peripheral pool from the thymus at rates ranging from 4×10^8 to 2×10^9 day^{-1} depending on age from 0 to 20 years, with a peak of 2×10^9 day^{-1} at about 1 year of age (15). In addition to thymic output, maintenance of the naive T cell pool in humans also depends on peripheral T cell division (13). Naive T

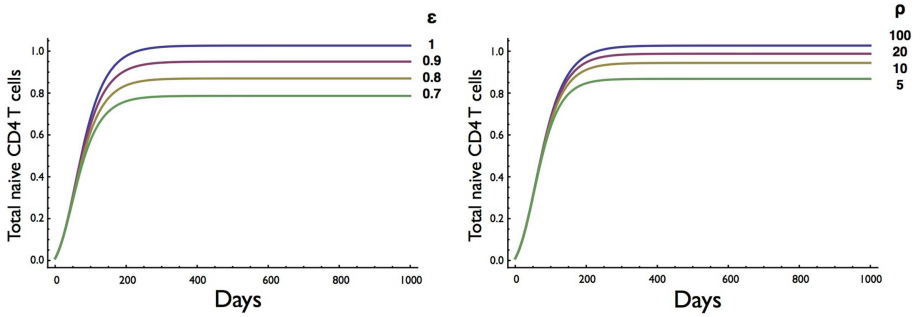


FIGURE 5 | Effect on T cell dynamics of changes in resource concentration for entry into cell division (ε) and survival (ρ). Other parameters are as in **Table 1**. It is noteworthy that the homeostatic equilibrium is more sensitive to changes in the resource parameter (ε) for entry into cell division than the parameter (ρ) for survival.

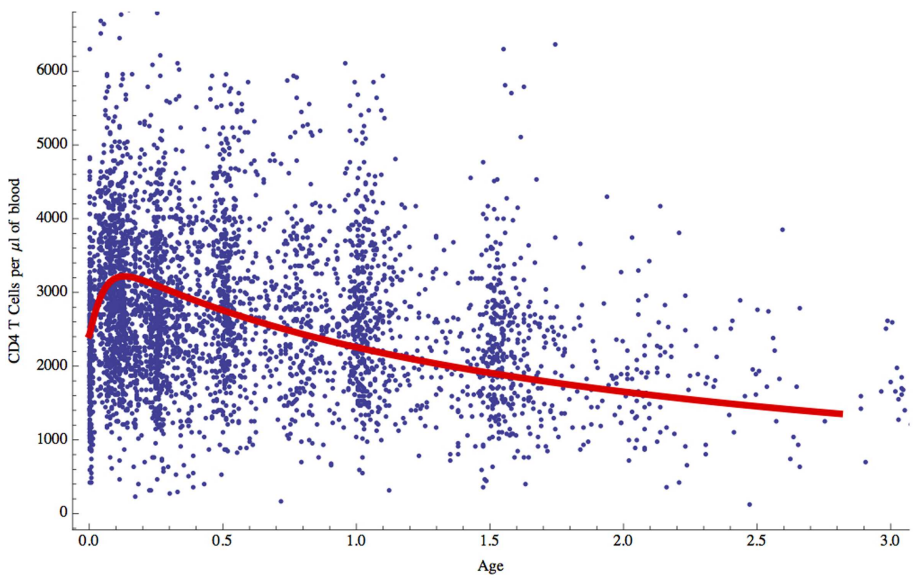


FIGURE 6 | Naive CD4 T cell concentrations (cells/μl of blood) predicted by the model for children aged 0–3 years (red curve) compared with clinical data for normal children.

cell entry into cell division occurs in response to TCR signaling by self-peptide MHC and signals provided by IL7 (4–9, 29). Competition for these resources determines the rate of entry (19, 29). Similarly, survival depends on signaling by IL7 albeit at a different concentration threshold than for proliferation (27). The rate of exit from the cell cycle was taken from the approximate time taken to complete cell division [6 h (19)] and the death rate of the cells in cycle was modeled by the described Fas/Fas ligand mechanism of activated cells (24).

The mathematical model described here consists of two coupled non-linear differential equations representing resting and dividing cell compartments with parameters for thymic export, entry into cell division, return to the resting state and death (at different rates) of resting and dividing cells. Exponential forms were used for entry into cell division and death of cells in the resting compartment to represent the known competition for resources for cell proliferation and survival. Alternative functional forms for density dependence may be worth exploring in the future. The model is a mechanistic model based on the known biology of naïve T cell homeostasis so that the different parameters all have a biological interpretation as indicated in the methods. An alternative but non-mechanistic mathematical model of T cell homeostasis has been described (30), which depends on assumptions about the inheritability of life spans and it cannot therefore be easily compared to the model we describe here. Rather, in our model proliferation and death rates depend on competition for resources as supported by experimental evidence. Our model does however assume the naïve T cell population is homogeneous without taking into account clonal diversity and it would be of interest in the future to develop stochastic ODEs or agent based models.

When T cell export from the thymus was kept constant to represent a young adult, simulated T cell numbers converged from either low or high initial levels to a stable homeostatic equilibrium consistent with cell numbers in a normal, healthy adult. This is concurrent with the increase in T cells observed in response to lymphopenia and the decrease following T cell expansion after infection (16, 17). Consistent with previous studies, the death rate of proliferating cells is higher than that of resting cells in our model (31). The death rate of resting cells found here also agrees approximately with experimental results (32–34). The average time between cell divisions was about 50 days compared to 60 days in the model described by Yates (31). Another interesting aspect of these results was the interdivision time of cells, calculated to be around 30 days. This is comparable to the results of deuterium labeling experiments, which suggest an average of 26 days (16, 17, 34, 35).

The results obtained by altering parameter values gave a clear indication of the effect of the different rates of cell death and proliferation. The corresponding expected rise and fall in the set point of the T cell pool was reassuring. This set point appeared to be more sensitive to increments in the death rate of resting cells than to increases of the same order in the activation rate. Importantly, the T cell numbers at equilibrium decreased as the resource term for entry into cell division (ε) or the resource term for rescue from cell death (ρ) decreased although the equilibrium was less sensitive to changes in the resource required for survival (**Figure 5**). The sensitivity of the homeostatic T cell equilibrium to a resource, such as IL7, is potentially important for understanding conditions resulting in reduced CD4 T cell numbers, such as HIV, and the degree of recovery after treatment with antiretroviral therapy (ART). In a recent study, the degree of CD4 T cell recovery in children on ART was correlated with the initial (pre-ART) CD4 T cell count and the length of time between infection (at birth) and the commencement of treatment (36). One explanation for this finding is that HIV infection compromises lymph node structure and hence the ability to provide resources required for homeostatic T cell division (37, 38). The two-compartment model could then be a valuable tool for exploring T cell homeostasis in HIV and other conditions such as T cell reconstitution following stem cell transplantation.

The other question addressed by the model was whether the incorporated biological mechanisms were in themselves sufficient to explain the known decrease in naive CD4 T cell concentration over the first few years of life when T cell export from the thymus increases to a maximum at 1 year of age and then declines, and the child is growing in size with an accompanying increase in blood volume (**Figure 1**). Total naive CD4 T cell numbers obtained from the model were converted into T cell concentration in the blood using blood volume/age data (14). The model's predictions were found to agree very well with real data from a cohort of children aged 0–3 years (**Figure 6**): the two-compartment model was able to reproduce the initial rise and subsequent slow decline in T cell count observed in healthy individuals over 0–3 years. These findings suggest that the changes in CD4 T cell counts in young children can be explained simply by the change in thymic output and body size as they grow and does not require any additional developmental changes to homeostatic mechanisms. It is important to point out that the thymic export model does not take memory cells into account. However, the proportion of memory cells in the CD4+ T cell pool in children is relatively small and therefore should not have a significant effect on these results (39).

In conclusion, we have presented a mechanistic two-compartment model of naive T cell homeostasis based on the known biology, which reproduces results obtained by other methods with good accuracy. It is likely to be an appropriate model for investigations of T cell reconstitution and homeostasis in diseases such as HIV, in patients given bone marrow transplantation and even for understanding reconstitution after thymic transplants for athymic patients with DiGeorge syndrome.

REFERENCES

1. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. *Science* (1999) **286**:958–61. doi:10.1126/science.286.5441.958
2. Freitas AA, Rocha B. Population biology of lymphocytes: the flight for survival. *Annu Rev Immunol* (2000) **18**:83–111. doi:10.1146/annurev.immunol.18.1.83
3. Jameson SC. T cell homeostasis: keeping useful T cells alive and live T cells useful. *Semin Immunol* (2005) **17**:231–7. doi:10.1016/j.smim.2005.02.003
4. Schluns KS, Kieper WC, Jameson SC, Lefrancois L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells in vivo. *Nat Immunol* (2000) **1**:426–32. doi:10.1038/80868
5. Tan JT, Dudl E, Leroy E, Murray R, Sprent J, Weinberg KI, et al. IL-7 is critical for homeostatic proliferation and survival of naive T cells. *Proc Natl Acad Sci USA* (2001) **98**:8732–7. doi:10.1073/pnas.161126098
6. Vivien L, Benoist C, Mathis D. T lymphocytes need IL-7 but not IL-4 or IL-6 to survive in vivo. *Int Immunol* (2001) **13**:763–8. doi:10.1093/intimm/13.6.763

7. Seddon B, Zamoyska R. TCR and IL-7 receptor signals can operate independently or synergize to promote lymphopenia-induced expansion of naive T cells. *J Immunol* (2002) **169**:3752–9.
8. Labrecque N, Whitfield LS, Obst R, Waltzinger C, Benoist C, Mathis D. How much TCR does a T cell need? *Immunity* (2001) **15**:71–82. doi:10.1016/S1074-7613(01)00170-4
9. Seddon B, Zamoyska R. TCR signals mediated by Src family kinases are essential for the survival of naive T cells. *J Immunol* (2002) **169**:2997–3005.
10. Saini M, Pearson C, Seddon B. Regulation of T cell-dendritic cell interactions by IL-7 governs T-cell activation and homeostasis. *Blood* (2009) **113**:5793–800. doi:10.1182/blood-2008-12-192252
11. Tough DF, Sprent J. Turnover of naive- and memory-phenotype T cells. *J Exp Med* (1994) **179**:1127–35. doi:10.1084/jem.179.4.1127
12. Asquith B, Borghans JA, Ganusov VV, Macallan DC. Lymphocyte kinetics in health and disease. *Trends Immunol* (2009) **30**:182–9. doi:10.1016/j.it.2009.07.013
13. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mogling R, Bregje De Boer A, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* (2012) **36**:288–97. doi:10.1016/j.immuni.2012.02.006
14. Bains I, Antia R, Callard R, Yates AJ. Quantifying the development of the peripheral naive CD4+ T cell pool in humans. *Blood* (2009) **113**:5480–7. doi:10.1182/blood-2008-10-184184
15. Bains I, Thiebaut R, Yates AJ, Callard RE. Quantifying thymic export: combining models of naive T cell proliferation and TREC dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**:4329–36. doi:10.4049/jimmunol.0900743
16. Yates AJ, Callard RE. Cell death and the maintenance of immunological memory. *Discrete Continuous Dyn Syst Ser B* (2001) **1**:43–60. doi:10.3934/dcdsb.2001.1.43
17. Yates AJ, Stark J, Klein N, Antia R, Callard RE. Understanding the slow depletion of memory CD4+ T cells in HIV infection. *PLoS Med* (2007) **4**:e177. doi:10.1371/journal.pmed.0040177
18. Steinmann GG, Klaus B, Muller-Hermelink HK. The involution of the ageing human thymic epithelium is independent of puberty. A morphometric study. *Scand J Immunol* (1985) **22**:563–75. doi:10.1111/j.1365-3083.1985.tb01916.x
19. Hogan T, Shuvaev A, Commenges D, Yates A, Callard R, Thiebaut R, et al. Clonally diverse T cell homeostasis is maintained by a common program of cell-cycle control. *J Immunol* (2013) **190**:3985–93. doi:10.4049/jimmunol.1203213
20. Seddon B, Tomlinson P, Zamoyska R. Interleukin 7 and T cell receptor signals regulate homeostasis of CD4 memory cells. *Nat Immunol* (2003) **4**:680–6. doi:10.1038/ni946
21. Takada K, Jameson SC. Naive T cell homeostasis: from awareness of space to a sense of place. *Nat Rev Immunol* (2009) **9**:823–32. doi:10.1038/nri2657
22. Yates AJ, Saini M, Mathiot A, Seddon B. Mathematical modelling reveals the biological programme regulating lymphopenia-induced proliferation. *J Immunol* (2008) **180**:1414–22.
23. Lynch DH, Ramsdell F, Alderson MR. Fas and FasL in the homeostatic regulation of immune responses. *Immunol Today* (1995) **16**:569–74. doi:10.1016/0167-5699(95)80079-4
24. Callard RE, Stark J, Yates AJ. Fratricide: a mechanism for T memory cell homeostasis. *Trends Immunol* (2003) **24**:370–5. doi:10.1016/S1471-4906(03)00164-9
25. Hellerstein M, Hanley MB, Cesar D, Siler S, Papageorgopoulos C, Wieder E, et al. Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. *Nat Med* (1999) **5**:83–9. doi:10.1038/4772
26. Vrisekoop N, Den Braber I, De Boer AB, Ruiter AF, Ackermans MT, Van Der Crabben SN, et al. Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool. *Proc Natl Acad Sci U S A* (2008) **105**:6115–20. doi:10.1073/pnas.0709713105
27. Palmer MJ, Mahajan VS, Chen J, Irvine DJ, Lauffenburger DA. Signaling thresholds govern heterogeneity in IL-7-receptor-mediated responses of naive CD8(+) T cells. *Immunol Cell Biol* (2011) **89**:581–94. doi:10.1038/icb.2011.5
28. Bunders M, Thorne C, Newell ML. Maternal and infant factors and lymphocyte, CD4 and CD8 cell counts in uninfected children of HIV-1-infected mothers. *AIDS* (2005) **19**:1071–9. doi:10.1097/01.aids.0000174454.63250.22
29. Kieper WC, Burghardt JT, Surh CD. A role for TCR affinity in regulating naive T cell homeostasis. *J Immunol* (2004) **172**:40–4.
30. Dowling MR, Hodgkin PD. Modelling naive T-cell homeostasis: consequences of heritable cellular lifespan during ageing. *Immunol Cell Biol* (2009) **87**:445–56. doi:10.1038/icb.2009.11
31. Yates AJ, Chan CCT, Callard RE. Modelling T cell activation, proliferation and homeostasis. In: Paton R, Mcnamara L, editors. *Multidisciplinary Approaches to Theory in Medicine*. Elsevier (2006). p. 281–308.
32. Ribeiro RM, Mohri H, Ho DD, Perelson AS. Modeling deuterated glucose labeling of T-lymphocytes. *Bull Math Biol* (2002) **64**:385–405. doi:10.1006/bulm.2001.0282
33. Macallan DC, Wallace D, Zhang Y, De Lara C, Worth AT, Ghattas H, et al. Rapid turnover of effector-memory CD4(+) T cells in healthy humans. *J Exp Med* (2004) **200**:255–60. doi:10.1084/jem.20040341
34. Asquith B, Zhang Y, Mosley AJ, De Lara CM, Wallace DL, Worth A, et al. In vivo T lymphocyte dynamics in humans and the impact of human T-lymphotropic virus 1 infection. *Proc Natl Acad Sci U S A* (2007) **104**:8035–40. doi:10.1073/pnas.0608832104
35. Ganusov VV, Borghans JA, De Boer RJ. Explicit kinetic heterogeneity: mathematical models for interpretation of deuterium labeling of heterogeneous cell populations. *PLoS Comput Biol* (2010) **6**:e1000666. doi:10.1371/journal.pcbi.1000666
36. Lewis J, Walker AS, Castro H, De Rossi A, Gibb DM, Giaquinto C, et al. Age and CD4 count at initiation of antiretroviral therapy in HIV-infected children: effects on long-term T-cell reconstitution. *J Infect Dis* (2012) **205**:548–56. doi:10.1093/infdis/jir787
37. Zeng M, Paiardini M, Engram JC, Beilman GJ, Chipman JG, Schacker TW, et al. Critical role of CD4 T cells in maintaining lymphoid tissue structure for immune cell homeostasis and reconstitution. *Blood* (2012) **120**:1856–67. doi:10.1182/blood-2012-03-418624
38. Zeng M, Southern PJ, Reilly CS, Beilman GJ, Chipman JG, Schacker TW, et al. Lymphoid tissue damage in HIV-1 infection depletes naive T cells and limits T cell reconstitution after antiretroviral therapy. *PLoS Pathog* (2012) **8**:e1002437. doi:10.1371/journal.ppat.1002437
39. Huenecke S, Behl M, Fadler C, Zimmermann SY, Bochennek K, Tramsen L, et al. Age-matched lymphocyte subpopulation reference values in childhood and adolescence: application of exponential regression analysis. *Eur J Haematol* (2008) **80**:532–9. doi:10.1111/j.1600-0609.2008.01052.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2013; accepted: 27 October 2013; published online: 11 November 2013.

*Citation: Hapuarachchi T, Lewis J and Callard RE (2013) A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. Front. Immunol. **4**:366. doi: 10.3389/fimmu.2013.00366*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Hapuarachchi, Lewis and Callard. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 December 2013
doi: 10.3389/fimmu.2013.00434

Mathematical model of naive T cell division and survival IL-7 thresholds

Joseph Reynolds[1], Mark Coles[2], Grant Lythe[1] and Carmen Molina-París[1]*

[1] *Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds, UK*
[2] *Centre for Immunology and Infection, University of York, York, UK*

Edited by:
Miles Davenport, University of New South Wales, Australia

Reviewed by:
Edward John Collins, The University of North Carolina at Chapel Hill, USA
Christian Schönbach, Nazarbayev University, Kazakhstan
Robin Callard, University College London, UK

***Correspondence:**
Carmen Molina-París, Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
e-mail: carmen@maths.leeds.ac.uk

We develop a mathematical model of the peripheral naive T cell population to study the change in human naive T cell numbers from birth to adulthood, incorporating thymic output and the availability of interleukin-7 (IL-7). The model is formulated as three ordinary differential equations: two describe T cell numbers, in a resting state and progressing through the cell cycle. The third is introduced to describe changes in IL-7 availability. Thymic output is a decreasing function of time, representative of the thymic atrophy observed in aging humans. Each T cell is assumed to possess two interleukin-7 receptor (IL-7R) signaling thresholds: a survival threshold and a second, higher, proliferation threshold. If the IL-7R signaling strength is below its survival threshold, a cell may undergo apoptosis. When the signaling strength is above the survival threshold, but below the proliferation threshold, the cell survives but does not divide. Signaling strength above the proliferation threshold enables entry into cell cycle. Assuming that individual cell thresholds are log-normally distributed, we derive population-average rates for apoptosis and entry into cell cycle. We have analyzed the adiabatic change in homeostasis as thymic output decreases. With a parameter set representative of a healthy individual, the model predicts a unique equilibrium number of T cells. In a parameter range representative of persistent viral or bacterial infection, where naive T cell cycle progression is impaired, a decrease in thymic output may result in the collapse of the naive T cell repertoire.

Keywords: IL-7, T cell, homeostasis, threshold, IL-7R, mathematical model, thymic output

1. INTRODUCTION

The number of naive T cells in the periphery is determined by a balance between cell loss (death or differentiation) and cell renewal due to cell division and thymic export (1, 2). In humans, at least, the decline in thymic export occurs mainly in childhood, from about a year of age until 20 years of age, when the number of naive T cells is increasing (3). In adults, the decline in thymic export is much less pronounced but the number of naive T cells is, more or less, constant (4). Survival of the naive T cell population in the periphery depends on both common gamma chain cytokines and weak "tonic" signals induced by recognition of self-peptides by the T cell receptor (TCR) (5, 6). IL-7 is required for the homeostatic expansion of naive $CD8^+$ and $CD4^+$ T cells in lymphopenic hosts, while naive T cells disappear over a 1-month period upon adoptive transfer into IL-7 deficient ($IL\text{-}7^-$) hosts (7–9).

Signals from recognition of self-peptides bound to major histocompatibility complex (sp-MHC), and IL-7, promote cell survival. Naive T cell survival is impaired when removing access to one of these signals (10–14). Of interest are the mechanisms by which these signals are regulated, and that result in a stable number of naive T cells throughout the lifetimes of mice and humans. In this paper, we focus on IL-7 as a master regulator of naive T cell survival (15). IL-7 is produced by stromal cells in tissues, including fibroblastic reticular cells, marginal reticular cells, and lymphatic endothelial cells (16). These cells produce very small amounts of IL-7 messenger RNA, consistent with IL-7 protein levels limiting T cell expansion. IL-7 is a heparin-sulfate binding protein, and as such, it will bind extra-cellular matrix surrounding stromal cells. Thus, the interaction between naive T cells and stroma controls their homeostasis (17). Recognition of higher affinity, non-self-peptides by the T cell receptor induces naive T cells to undergo an alternative, IL-7 independent, survival program dependent on IL-2 (18).

Naive $CD8^+$ T cell responses depend on the amount of IL-7 cells are exposed to (19). At low IL-7 concentrations ($<10^{-2}$ ng ml^{-1}), cell viability was impaired; at higher concentrations (>1 ng ml^{-1}) cells were observed to proliferate in response to IL-7. This difference might arise from changes in the strength of the IL-7R induced signal the cell receives. For an individual cell, IL-7R induced signaling must be greater than some threshold to prevent the accumulation of pro-apoptotic proteins. Similarly, IL-7R signaling must be greater than a second, higher, threshold to induce cell division. Heterogeneity at the single cell level in IL-7 signaling thresholds (a property reported to depend on expression of IL-7R), resulted in differential survival and division (19). Although these observations are based on two different $CD8^+$ T cell receptor transgenic mice, it is assumed that the key principles regarding T cell survival will be found in the repertoire of naive $CD4^+$ and $CD8^+$ T cells.

We introduce a deterministic mathematical model of the naive T cell population to study the change in human naive T cell numbers from birth to adulthood. We will assume cell survival

depends on the availability of IL-7. We do not include availability of sp-MHC as a variable within the model, but assume sp-MHC availability is sufficient to allow cell survival and proliferation, in conjunction with sufficient IL-7 stimulus. We also make the approximation that heterogeneity is constant with changes in age. For a mathematical study of the impact sp-MHC availability has on clonal diversity, the reader is referred to Stirk et al. (20, 21). Our model is a mathematical description of the homeostasis of the naive T cell repertoire, but does not consider stimulation by foreign antigens.

2. MATERIALS AND METHODS

2.1. A MATHEMATICAL DESCRIPTION OF THE SIZE OF THE PERIPHERAL NAIVE T CELL POPULATION

Stochastic processes provide a method of treating each cell as a distinct, countable object, and permit a more realistic model than a deterministic characterization. Fluctuations in the number of cells can be considered but, in a non-linear stochastic model, approximations are often made to facilitate the analysis. In the linear noise approximation (22), for example, fluctuations are assumed to be of order $\Omega^{\frac{1}{2}}$ for a system of size Ω. The human peripheral T cell compartment is estimated to contain of the order of 10^{11} T cells (3). Letting the system size be the average number of naive T cells in humans, we find $\mathcal{O}(\Omega) = 10^{11}$ cells, and correspondingly, fluctuations are expected to be typically $10^5 - 10^6$ cells in magnitude. That is, we expect fluctuations of approximately 0.001% in the size of the human naive T cell pool due to stochasticity in the per cell division and death rates. Based on these considerations, adopting a deterministic approach to describe the total human peripheral naive T cell population is reasonable.

We assume peripheral naive T cells are either in a resting state, or proceeding through the cell cycle. The deterministic variables $R(t)$ and $C(t)$ are introduced to model the total number of T cells in the resting and cycling states, respectively. The variable $I(t)$ is introduced to model the concentration of IL-7. The deterministic approach we take does not consider any notion of space. Indeed, this approach is tantamount to assuming the resource, IL-7, is shared equally amongst all cells. Competition for the resource is introduced only so far as each cell acts to reduce the global concentration of the resource. Resting cells may receive a signal which induces them to proceed through one round of division. Upon completion of the cell cycle, a cycling cell produces two daughter cells in the resting compartment. Resting cells are assumed to die if the IL-7 induced survival signal is insufficient; cells may also die during cell cycle. The input of cells from the thymus into the resting compartment, in keeping with observations in humans, is a decreasing function of time (23, 24). Production of IL-7 is related to the size of the lymphatic system architecture, which we estimate from the body mass of an individual. In the absence of T cells, IL-7 is assumed to be degraded and/or consumed by other cell types at a constant rate. Upon signal induction through the IL-7 receptor, IL-7 is assumed to be consumed by the T cell. A diagrammatic representation of the model is given in **Figure 1**.

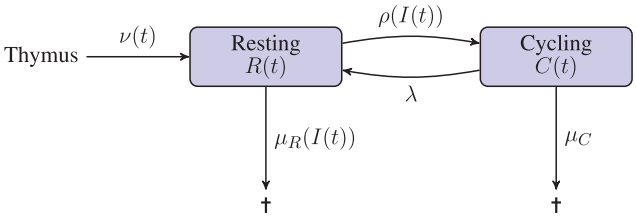


FIGURE 1 | Diagrammatic illustration of the deterministic model. T cells leaving the thymus enter the resting naive peripheral pool. Cells in either a resting or cycling state may die. The rate of death from the resting state depends on the availability of the resource (IL-7), whereas the death rate for cycling cells is constant. Resting cells enter the cell cycle at a rate that depends on the availability of the resource (IL-7). Cycling cells produce two daughter cells in the resting state upon completion of the cell cycle.

2.2. IL-7 SIGNALING AND HETEROGENEITY IN IL-7 THRESHOLDS

In the model, IL-7 signaling is assumed to be uniform across the population. Yet, we introduce heterogeneity in the signaling thresholds for survival and proliferation. Let $S(t)$ be the average signaling strength across the naive T cell population. Each cell experiences the same strength of signaling for a given concentration of IL-7, $I(t)$. We relate the internal signaling to the concentration of IL-7 by the equation

$$S(t) = \frac{600I(t)}{0.025 + I(t)}. \tag{1}$$

The functional form and constants of this relationship are derived from the study of IL-7 receptor dynamics summarized in the Appendix. We assume each individual cell possesses a unique pair of thresholds for survival and proliferation. Furthermore, we assume, in the continuous limit, that these thresholds are distributed log-normally across the entire population of cells. The use of log-normal distributions guarantees, first of all, that all signaling thresholds are positive real numbers. Secondly, the log-normal distribution ensures that no cell can survive or divide independently of IL-7.

Let the random variable Θ_s represent the survival threshold, and let Θ_p represent the proliferation threshold. We write

$$\Theta_s \sim \log\mathcal{N}\left(\log\theta_s, \frac{1}{2\alpha^2}\right), \quad \Theta_p \sim \log\mathcal{N}\left(\log\theta_p, \frac{1}{2\alpha^2}\right), \quad \alpha \in \mathbb{R}^+. \tag{2}$$

The respective probability density functions are

$$p_{\Theta_x}(\theta) = \frac{\alpha}{\sqrt{\pi}\theta}\exp\left[-\left(\alpha\left(\log\theta - \log\theta_x\right)\right)^2\right], \quad x = s,\ p. \tag{3}$$

Estimates for θ_s and θ_p are obtained from the model summarized in the Appendix. Lauffenburger et al. found a significant change in cell viability for both OT-1 and F-5 T cells at around $10^{-2.5}$ ng ml^{-1} IL-7 (19). Proliferation of OT-1 cells occurred above 1 ng ml^{-1} IL-7. We estimate the equilibrium signaling at these concentrations as 60 and 600 units, respectively, setting $\theta_s = 60$ and $\theta_p = 600$. Modeling heterogeneity in IL-7 responses by assuming heterogeneity in the IL-7 signaling thresholds, allows us to avoid modeling the naive population using either: (i) a PDE

approach, where heterogeneity is continuous across the population of cells, or (ii) describing each subset of cells sharing common thresholds with its unique governing set of ODEs. However, these approaches present an obvious avenue for further research beyond the scope of this paper.

2.2.1. Death rate of resting cells

Suppose each T cell in the naive population is a distinct member possessing a unique signaling threshold for survival. As in the previous section, we assume these signaling thresholds are distributed log-normally (in the continuous limit). We suppose the death rate of an individual cell is Boolean, in the sense that if the global signaling strength is greater than the cell's individual survival threshold, then the cell can survive indefinitely. Similarly, if the signaling strength is below the survival threshold, the cell will undergo apoptosis at a rate μ_R. The death rate for cell i with survival threshold $\theta_s^{(i)}$, is given by

$$f_s\left(S(t),\, \theta_s^{(i)}\right) = \begin{cases} \mu_R & \text{if } S(t) < \theta_s^{(i)}, \\ 0 & \text{if } S(t) \geq \theta_s^{(i)}. \end{cases} \tag{4}$$

In the continuous limit (assuming signaling thresholds are distributed log-normally), the average death rate for the population of naive T cells is given by

$$\begin{aligned} \bar{\mu}_R(S(t)) &= \int_0^\infty f_s(S(t),\, \theta)\, p_{\Theta_s}(\theta)\, \mathrm{d}\theta = \int_{S(t)}^\infty \mu_R p_{\Theta_s}(\theta)\, \mathrm{d}\theta, \\ &= \frac{1}{2}\mu_R\left[1 - \operatorname{erf}\left(\alpha\left(\log S(t) - \log\theta_s\right)\right)\right], \end{aligned} \tag{5}$$

where $p_{\Theta_s}(\theta)$ is the probability density function of the random variable Θ_s, defined by equation (3) with $x = s$.

2.2.2. Rate of entry into cell cycle

Analogous to Section 2.2.1, we assume each T cell in the naive population is a distinct member possessing a unique signaling threshold for proliferation. We let the individual rate of entry into cell cycle be given by

$$f_p\left(S(t),\, \theta_p^{(i)}\right) = \begin{cases} 0 & \text{if } S(t) < \theta_p^{(i)}, \\ \rho & \text{if } S(t) \geq \theta_p^{(i)}. \end{cases} \tag{6}$$

Assume, in the continuous limit, the signaling threshold for entry into the cell cycle is represented by the random variable Θ_p, defined in equation (2), with probability density function $p_{\Theta_p}(\theta)$ (equation (3), $x = p$). The average rate of entry into cell cycle is given by

$$\begin{aligned} \bar{\rho}(S(t)) &= \int_0^\infty f_p(S(t),\, \theta)\, p_{\Theta_p}(\theta)\, \mathrm{d}\theta = \int_0^{S(t)} \rho p_{\Theta_p}(\theta)\, \mathrm{d}\theta, \\ &= \frac{1}{2}\rho\left[1 + \operatorname{erf}\left(\alpha\left(\log S(t) - \log\theta_p\right)\right)\right]. \end{aligned} \tag{7}$$

2.2.3. Cell cycle progression

Cycling cells take on average λ^{-1} days to complete the cell cycle. After a cell divides, both daughter cells are produced in the resting state and require a second signal before they can progress through another round of cell division. Cell cycle may be interrupted resulting in the death of the cell. Such death events occur at a rate μ_C.

2.2.4. Thymic export

We assume thymic output to be a decreasing function of time. In particular, we use the functional form given by Bains et al. (3). Let us introduce the thymic output function, $\nu(t)$, as follows

$$\begin{aligned} \nu(t) = {} & 2.32 \times 10^8 \exp\left(-1.1 \times 10^4 t\right) \\ & + 1.15 \times 10^8 \exp\left(-1.6 \times 10^7 t^2\right), \end{aligned} \tag{8}$$

where t corresponds to the age of the individual, measured in days. A plot of this function is shown in the left panel of **Figure 2**. The function was chosen by Bains et al. to describe the rate of thymic export of $CD4^+$ T cells. We use the same function to describe the export rate of all naive T cells ($CD4^+$ or $CD8^+$ T cells). This approximation is justified since we require the absolute cell count to roughly approximate the cell count observed in humans (indeed, such an observation is likely subject to large differences). Of interest later in the paper is the relative variation of cell numbers with different choices of parameter values. For our purposes, the important feature of $\nu(t)$ is that it is a decreasing function of time.

2.2.5. Internalization of IL-7

We use the model summarized in the Appendix to estimate the rate of IL-7 internalization. The total number of IL-7 molecules internalized by a single T cell in 1 day, exposed to IL-7 at concentration I ng ml^{-1}, is described by the function

$$\frac{6.7I(t)}{2 + I(t)}\left(3 + \frac{5}{I(t) + 3 \times 10^{-2}}\right) \times 10^3 \text{ cell}^{-1} \text{ day}^{-1}. \tag{9}$$

It is reported IL-7 has a molecular mass of 17 kDa ($\approx 2.8 \times 10^{-11}$ ng) (25). Based on this, we define the per cell rate of IL-7 internalization to be

$$\frac{2I(t)}{2 + I(t)}\left(3 + \frac{5}{I(t) + 3 \times 10^{-2}}\right) \times 10^{-7} \text{ ng day}^{-1} \text{ cell}^{-1}.$$

In order to convert the rate of change of mass to the rate of change of concentration, we must choose a volume for the system. Naive T cells are typically found in the lymph nodes, spleen, and gut of the human body. We shall make the rough estimation that the total volume is of the order of 1 l. This implies the rate of IL-7 internalization by all naive T cells in the population is given by

$$\begin{aligned} \gamma(I(t))\, R(t) = {} & \frac{2I(t)}{2 + I(t)}\left(3 + \frac{5}{I(t) + 3 \times 10^{-2}}\right) \\ & \times 10^{-10} R(t) \text{ ng ml}^{-1} \text{ day}^{-1}, \end{aligned} \tag{10}$$

where $R(t)$ is the number of resting naive T cells. The IL-7 internalization rate, $\gamma(I(t))$, is shown in the middle panel of **Figure 2**.

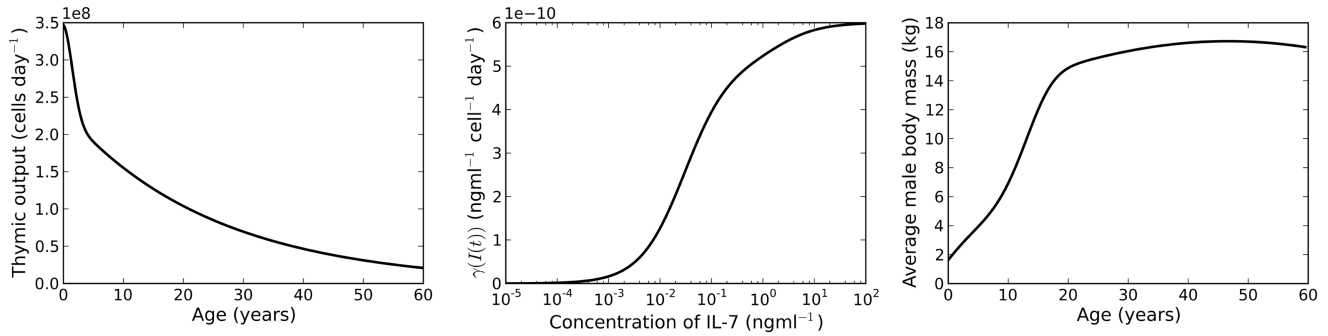


FIGURE 2 | Left panel: rate of thymic export. Plot of equation (8). Middle panel: IL7 internalization rate. Plot of the function $\gamma(I(t))$. Right panel: plot of the function $m(s)$ (equation (11)), a model for the average body mass of male individuals.

2.2.6. Production of IL-7

We assume the rate of IL-7 production is proportional (with proportionality constant $\hat{\beta}$) to the body mass of an individual. To estimate average body mass we use the model given by Burmaster and Crouch (26). The explicit relationship between mass and age is given by the function $m(s)$ as follows

$$m(s) = \exp\left[4.1 + 1.4 \times 10^{-2}s - 1.5 \times 10^{-4}s^2 - 2.0\exp\left[-s\left(0.15 - 1.4 \times 10^{-2}s + 9.8 \times 10^{-4}s^2\right)\right]\right], \tag{11}$$

where s is age measured in years. A plot of this function is given in the right panel of **Figure 2**. The rate of IL-7 production is given by the function

$$\beta(t) = \hat{\beta}\, m(365t), \tag{12}$$

where t denotes age as measured in days.

2.2.7. Intra-cellular degradation of IL-7

We assume IL-7 is degraded and internalized by other cell types at a constant rate. We let the degradation rate of IL-7 be denoted by the parameter δ.

2.2.8. Deterministic mathematical model

From the above assumptions, the system of differential equations governing the behavior of the naive T cells (resting and cycling) and the concentration of IL-7 is given by

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta(t) - \gamma(I(t))\,R(t) - \delta I(t), \tag{13}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \nu(t) - \left[\bar{\rho}(S(t)) + \bar{\mu}_R(S(t))\right]R(t) + 2\lambda C(t), \tag{14}$$

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = \bar{\rho}(S(t))\,R(t) - (\mu_C + \lambda)\,C(t). \tag{15}$$

This system is subject to the initial conditions I_0, R_0, and C_0.

3. RESULTS

3.1. PARAMETER ESTIMATES

The function describing the rate of internalization of IL-7, equation (10), and the signaling relation, equation (1), were both estimated from our studies of IL-7 receptor dynamics in naive T cells (summarized in the Appendix). In the same study we have estimated the average signaling thresholds to be, respectively, $\theta_s = 60$ and $\theta_p = 600$. Our estimate for θ_p was found to be close ($\theta_p \approx 585$, which we have rounded to 600) to the limit of the signaling function equation (1), as $I(t) \to +\infty$. These estimates imply that, on average, approximately half of the naive T cell population does not proliferate in response to excess IL-7 when receptor dynamics is in equilibrium. Such a finding is consistent with the observations made by Lauffenburger et al. in the experiments described in Ref. (19). In these experiments, F-5 T cells did not proliferate (whereas OT-1 cells did proliferate), even in excess amounts of IL-7. Estimates for the IL-7 average signaling thresholds, θ_s and θ_p, have been based on OT-1 cells. However, we note that changes to these thresholds only result in quantitative differences, provided $\theta_s < \theta_p$.

We assume cycling naive T cells take 12 h to complete the cell cycle and produce two daughter cells ($\lambda^{-1} = 0.5$ day). Resting naive T cells are assumed to die after 2 days of IL-7 starvation $\left(\mu_R^{-1} = 2\ \text{day}\right)$. We, furthermore, set the rate of entry into cell cycle to be $\rho = 5\ \text{day}^{-1}$. Cycling cells are assumed to die at rate $\mu_C = 1\ \text{day}^{-1}$ in healthy individuals, whereas we choose $\mu_C = 3\ \text{day}^{-1}$ ($>\lambda$) to be representative of cell cycle impairment. These parameters have been estimated from the literature. We note that the model behavior we discuss in the following sections was found to be robust to changes in these parameters. Robustness was concluded since a 10-fold change in each or any combination thereof, of these four parameters did not change the qualitative behavior of the model, provided the relation $\lambda > \mu_C$ (or $\lambda < \mu_C$) was maintained. We examine the changes in model behavior arising from altering the relative values of λ and μ_C in the following sections.

We choose δ such that $\delta I(t)$ is similar in magnitude to $\gamma(I(t))R(t)$, when $R(t) \approx 10^{11}$, and $I(t) \approx 10^{-2}\,\text{ng ml}^{-1}$. The proportionality constant $\hat{\beta}$ is chosen such that we observe $\mathcal{O}(10^{11})$ naive T cells in equilibrium at 20 years of age (for $I(t) \approx 10^{-2}\,\text{ng ml}^{-1}$). We had the least inclination when choosing

Table 1 | Parameter choices for the mathematical model.

Parameter	Value	Units
θ_s	60	Signaling units
θ_p	600	Signaling units
α	2	(Log signaling units)$^{-1}$
$\hat{\beta}$	0.2	10^{-12} ml^{-1} day^{-1}
δ	500	day^{-1}
λ	2	day^{-1}
ρ	5	day^{-1}
μ_R	0.5	day^{-1}
μ_C	1	day^{-1}

the parameter α. This parameter describes the spread in the individual signaling thresholds across the naive T cell population. We choose $\alpha = 2$, however this choice has no justification from the literature. The parameter set is summarized in **Table 1**.

3.2. THERE EXISTS A UNIQUE AND STEADY STATE WHEN $\lambda > \mu_C$

Let us suppose changes in thymic output and IL-7 production occur in time scales slower than those of the changes in the number of naive T cells. Under this assumption, we look for adiabatic solutions of the system as follows:

$$0 = \beta(t) - \gamma\left(\hat{I}(t)\right)\hat{R}(t) - \delta\hat{I}(t)\ , \tag{16}$$

$$0 = \nu(t) - \left(\bar{\rho}\left(\hat{S}(t)\right) + \bar{\mu}_R\left(\hat{S}(t)\right)\right)\hat{R}(t) + 2\lambda\hat{C}(t)\ , \tag{17}$$

$$0 = \bar{\rho}\left(\hat{S}(t)\right)\hat{R}(t) - (\mu_C + \lambda)\hat{C}(t)\ . \tag{18}$$

For the parameter set studied, the relative error between this solution and the exact solution is within 3% for the resting naive T cell population and the concentration of IL-7, and within 14% for the cycling T cell population (see **Figure 3**).

All numerical results presented in the paper have been obtained with a Python code[1]: differential equations (13), (14), and (15) have been solved using a fourth-order Runge-Kutta scheme. Quasi-stationary solutions were found using the scipy.optimize package. Bifurcation plots were computed using a bisection scheme to search for multiple solutions of equation (21) in the interval [0,1]. Corresponding T cell numbers were calculated using equations (19) and (20).

Notice that the adiabatic solution for both cell types is uniquely defined for a given value of cytokine concentration, $\hat{I}(t)$, namely

$$\hat{R}(t) = \frac{\beta(t) - \delta\hat{I}(t)}{\gamma\left(\hat{I}(t)\right)}\ , \tag{19}$$

$$\hat{C}(t) = \frac{\bar{\rho}\left(\hat{S}(t)\right)}{\mu_C + \lambda}\frac{\beta(t) - \delta\hat{I}(t)}{\gamma\left(\hat{I}(t)\right)}\ . \tag{20}$$

[1]Python code available upon request.

The problem of finding adiabatic solutions can then be reduced to finding solutions, $\hat{I}(t)$, to the one-dimensional equation

$$\frac{\nu(t)\,\gamma\left(\hat{I}(t)\right)}{\beta(t) - \delta\hat{I}(t)} = \bar{\mu}_R\left(\hat{S}(t)\right) + \left(1 - \frac{2\lambda}{\lambda + \mu_C}\right)\bar{\rho}\left(\hat{S}(t)\right)\ . \tag{21}$$

By construction, $\gamma(\hat{I}(t))$ is a monotonically increasing function of $\hat{I}(t)$. Furthermore, the existence of positive adiabatic solutions requires $\beta(t) > \delta\hat{I}(t)$. Therefore, for a fixed time t, the left-hand side of equation (21) is an increasing function of $\hat{I}(t)$. For $\lambda > \mu_C$, the right-hand side of equation (21) is a monotonically decreasing function of $\hat{I}(t)$. Lastly, $\gamma(\hat{I}(t)) = 0$ for $\hat{I}(t) = 0$, and the limit as $\hat{I}(t) \to 0$ of the right-hand side is equal to μ_R. It follows that the intersection of the left and right sides must be unique and positive. We deduce that for $\lambda > \mu_C$ there exists a unique adiabatic solution to the system (see left plot of **Figure 4**). This solution is stable for the parameter set given in **Table 1**. We have also numerically explored parameter space, but have not found a parameter set for which this solution is unstable. In **Figure 5** we present numerical solutions to equations (13–15), computed with a fourth-order Runge-Kutta method implemented in Python. Initial conditions were chosen to be the adiabatic solutions of equations (16–18) at $t = 0$.

3.3. THERE EXIST TWO STEADY STATES WHEN $\lambda < \mu_C$

Let us now consider the case $\lambda < \mu_C$. The right-hand side of equation (21) is no longer a decreasing function of $\hat{I}(t)$. We find there may exist up to three solutions $\hat{I}(t)$, two of which may be stable simultaneously, whilst the third is unstable (see right plot of **Figure 4**). Further examination (by numerically finding all solutions using the bisection method) of the model reveals a saddle-node bifurcation as thymic output changes with age (see left panel, **Figure 6**). We assume that individuals with (around) 10^{11} naive T cells are healthy, in the sense that they have a sufficient number of T cells to provide protection against immune challenges. When a cell in cycle is more likely to die than to produce two daughter cells, we define cell cycle progression to be impaired. Mathematically, this corresponds to $\lambda < \mu_C$. When cell cycle progression is impaired, the model predicts an individual may possess a healthy number of T cells, thereby being immuno-competent, up until the age at which the model bifurcates. For the parameter set we have investigated, this bifurcation is inevitable given the estimated decline in thymic output established by Bains et al. (3). Indeed, for a given parameter set, from the known rate at which thymic output declines, one can estimate the age at which the model bifurcates. The bifurcation results in a decrease in the naive T cell population size of approximately two orders of magnitude, that is, following the bifurcation we expect roughly 99 out of every 100 naive T cells to be lost.

4. DISCUSSION

In a healthy individual, it is reasonable to expect that naive T cells entering the cell cycle are more likely to complete division and produce two daughter cells, than to die during the division process. Therefore, we suppose the parameter relation $\lambda < \mu_C$ represents a

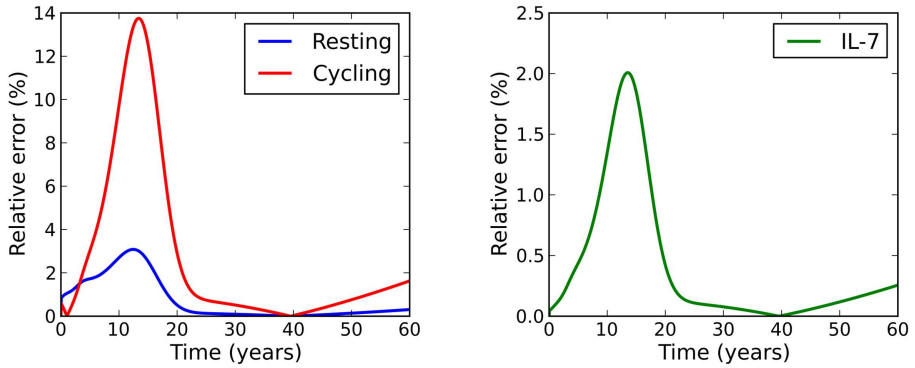


FIGURE 3 | Relative error between the adiabatic solution and the exact solution, where initial conditions were chosen to be equal to the adiabatic solution at time $t = 0$.

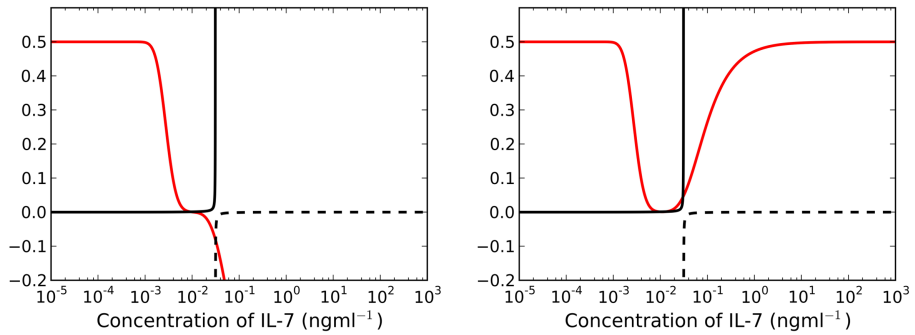


FIGURE 4 | Left panel: for $\lambda > \mu_C$ ($t = 25$ years), the right-hand side of equation (21) is a decreasing function of $I(t)$ (red line). There exists a unique, asymptotically stable solution $\hat{I}(t)$ found at the intersection of the solid black and red lines. Right panel: there exist three intersections between the red and solid black lines when $\lambda < \mu_C$ for $t = 25$ years. We require $\beta(t) > \gamma(\hat{I}(t))$ for existence of stable solutions, therefore we neglect all intersections with the dashed black line.

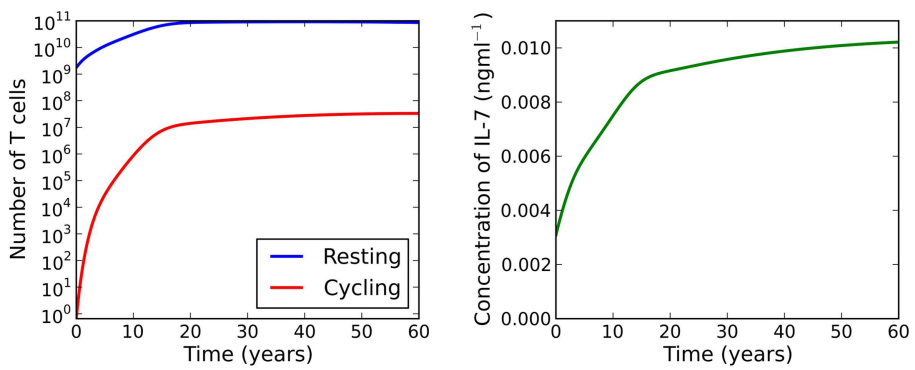


FIGURE 5 | Numerical solutions to equations (13–15) for $\lambda > \mu_C$. Initial conditions are chosen to be the adiabatic solutions of equations (16–18) at $t = 0$. We observe a large increase in the naive T cell population from birth to adulthood. During this period we see a 10^7-fold increase in the number of cycling cells, resulting from increased IL-7 availability and reduced thymic output. IL-7 availability increases continually as the individual ages.

healthy individual. We have shown, under this hypothesis, that the model allows a single asymptotically stable adiabatic solution. The total number of naive T cells per kilogram of body mass was found to increase over the first 18 years of life and decrease thereafter. The decline in this ratio for adults is seemingly a consequence of the reduction in thymic output.

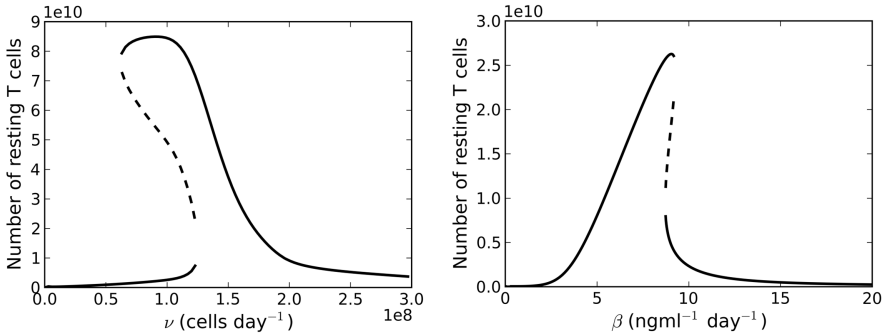


FIGURE 6 | Left panel: bifurcation diagram for varying the parameter $\nu(t)$ for $\lambda < \mu_C$. Right panel: bifurcation plot as a function of varying the rate of IL-7 production. Thymic export is fixed at ν(60 years).

The number of cycling naive T cells in adiabatic conditions was found to increase from practically 0 (approximately 0.75 cells) cells at birth to 10^7 cells by adulthood. Whilst the number of cycling cells increased thereafter, the increase was at a markedly slower rate. The increase in the number of cycling cells is probably due to the decline in thymic output, wherein competition for IL-7 decreases. Despite the increase in the cycling population, the number of cells in cycle is several orders of magnitude smaller than the resting population. For adults, this proportion is approximately $[0.01, 0.04]$% of the total naive population.

When cell cycle is impaired ($\lambda < \mu_C$), the model exhibits a saddle-node bifurcation as thymic output declines. For the parameter set investigated, the two adiabatic solutions are generally separated by two orders of magnitude. This feature motivates the following theoretical scenario: suppose a healthy individual experiences some event which causes naive cell cycle progression to become impaired at 18 years of age. The model predicts the naive T cell repertoire will experience a dramatic reduction in cell numbers at roughly 33 years of age. See **Figure 7** for the full solution of the model illustrating this scenario. From an immunological perspective, this scenario is interesting because the noticeable effect of the event (the dramatic loss of naive T cells) occurs roughly 15 years later than the event itself (naive cell cycle progression to become impaired at 18 years of age). Indeed, it is the reduction in thymic output that triggers the loss rather than the event itself. If the thymus was not atrophic, the loss in naive T cells would not occur.

Suppose now we fix thymic output at a constant rate and let the cell cycle be impaired. The model undergoes a saddle-node bifurcation as the rate of IL-7 production changes. More specifically, the model bifurcates, resulting in a decrease in the total number of T cells, as a consequence of increasing IL-7 production. Intuitively, this can be understood as follows: for increased IL-7 production, the rate of entry into cell cycle is enhanced, however, since cells are more likely to die in cell cycle than to produce two daughter cells, the enhanced rate of entry into cell cycle actually serves to decrease the total amount of naive T cells. The bifurcation diagram for this scenario is shown in the right panel of **Figure 6**. We found the critical value of this bifurcation decreases with thymic output. In the limiting case, corresponding to thymic output at 60 years of age, we found this critical value to be approximately 9.1 ng ml^{-1}. Consider again the theoretical scenario in which a healthy individual undergoes an event resulting in cell cycle impairment at 18 years of age. Suppose now at age 25 the same individual undergoes some treatment to limit IL-7 production to 8 ng ml^{-1}, corresponding to thymic production at approximately 11 years of age. The T cell count is reduced by approximately 60%. This reduction is a significant improvement on the 99% T cell loss in the untreated individual at 33 years of age. For this theoretical scenario, the model predicts limiting IL-7 availability will partially avoid the dramatic T cell loss arising from reducing thymic export when the cell cycle is impaired. See **Figure 8** for the full model solution in the treated individual.

In the model presented here we have neglected the fact naive T cells become activated in response to recognition of ligand specific to their unique TCR. Research by Koenen et al. has shown T cell survival is IL-7 independent following T cell activation (18). Suppose now we include a term in the governing ODEs to represent differentiation of naive T cells into cells with a different phenotype, such as activated T cells. Assuming no reversion back to the naive phenotype, such a term would appear in the model as a loss term equivalent to the death rate $\hat{\mu}_R(S(t))$. The simplest approach to including differentiation (due to activation) would be to assume naive T cells differentiate into activated T cells at a constant rate proportional to the rate of antigenic challenge. In this case, we would replace the term $\hat{\mu}_R(S(t))$ by $\hat{\mu}_R(S(t)) + \mu_D$, where μ_D is constant. Consider again the red curves in **Figure 4**. The differentiation term will cause a translation in the red curve of length μ_D up the vertical axis. For $\lambda > \mu_C$, there still exists a unique solution, however there will be a quantitative change in comparison to the case when $\mu_D = 0$. When $\lambda < \mu_C$, there still exists the possibility of a saddle-node bifurcation for a general parameter set. For the parameter set we have investigated, there exists a maximum value μ_D^*, for a given time t, such that we can find more than one solution. For differentiation rates $\mu_D > \mu_D^*$, we can only find a unique solution corresponding to the stable adiabatic one, in which we have reduced T cell numbers.

In this paper we have developed and analyzed a deterministic mathematical model of a population of naive T cells, whose survival depends on the availability of the cytokine IL-7. We have shown this model predicts a stable population of cells when cell cycle progression is healthy. More interestingly, when cell cycle progression is impaired, our results indicate declining thymic

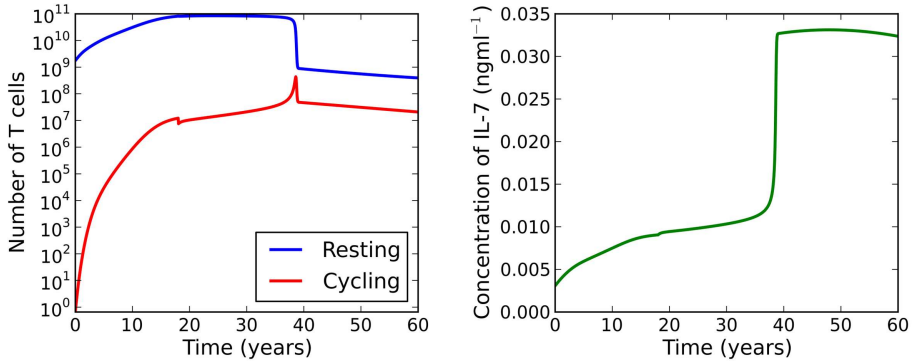


FIGURE 7 | Theoretical scenario in which we set $\mu_c = 3$ at 18 years of age. At roughly 33 years of age the model bifurcates resulting in a dramatic loss of naive T cells.

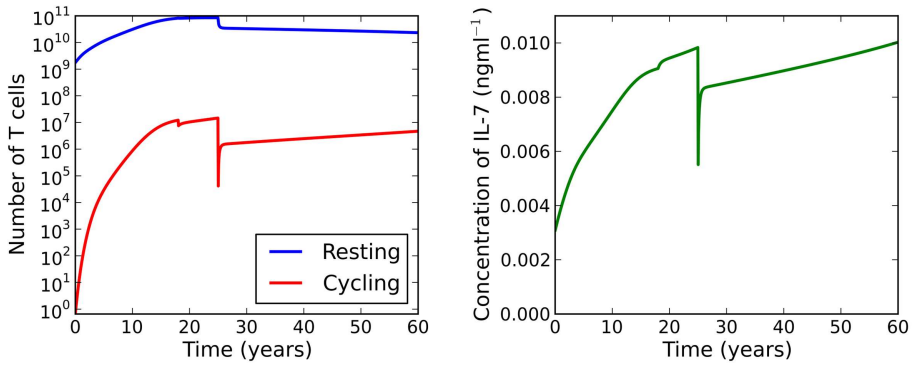


FIGURE 8 | At age 18 the individual experiences some event resulting in impaired cell cycle progression. At age 25 the individual undergoes treatment to limit IL-7 production to levels comparable to those at 11 years of age. Whilst the T cell count is reduced, this attrition is a significant improvement on the reduction observed when the individual has not received treatment (see **Figure 7**).

output may result in a dramatic loss in the number of naive T cells. Furthermore, we have been able to establish that limiting IL-7 production partially rescues this decline. However, our study has been restricted to naive T cells and we have not taken into account the accumulation of memory T cells as an individual ages. Memory T cells are generated in response to antigen or homeostatic cytokines (27), and during repeated homeostasis-driven divisions of naive T cells (28). Previous studies have shown the percentage of memory T cells increases with age, yet the percentage of naive T cells decreases with age (29). $CD4^+$ memory T cells also require IL-7 for survival and proliferation in the periphery (30). In the case of $CD8^+$ memory T cells, IL-15 is largely responsible for their peripheral survival, but IL-7 may also be required. It is, then, reasonable to assume that memory T cells compete for the IL-7 required for naive T cell survival. At least, that is, for those memory T cells which access IL-7 in the same tissues as naive T cells, such as the lymph nodes. As mentioned before, memory T cells can be derived from repeated divisions of naive T cells (31). In our model, the naive T cells that will acquire a memory-like phenotype first are those with low IL-7 survival and proliferation thresholds. We would expect, then, that over time the ratio of naive to memory T cells would decrease. Furthermore, the naive population would lose those cells with lowest survival and division thresholds. Given the distribution of signaling thresholds in our model, we would expect to see a shift to the right for the distribution of both survival and signaling thresholds. This shift, which implies that on average naive T cells require a higher concentration of IL-7 to enter cell cycle with age, together with the additional competition from an increasing memory population, is expected to cause a decrease in the total number of naive T cells and the percentage of naive T cells in cell cycle. These considerations might help explain the discrepancy between our model and data showing a reduction in the percentage of naive T cells expressing $Ki67^+$ during childhood[2]. There is support for the fact that childhood might be the phase in which most memory T cells are acquired (3). Indeed, it might be that the model presented here, better describes the total number of T cells that require IL-7, naive and memory, rather than just the naive T cell pool alone. This, however, would require a more detailed analysis, out of the scope of this paper.

[2]Our current model suggests the percentage of naive T cells in cell cycle is increasing.

Many different receptor-mediated signals are integrated by T cells in their micro-environment, such as cytokines, adhesion molecules, T cell receptors, and co-receptors (CD28, CTLA-4). However, the survival of naive T cells, the focus of this paper, has been shown to depend on IL-7 and low level TCR stimulation (8, 9, 15, 32). Other cytokines, such as IL-15, play a significant role in the homeostasis of memory $CD8^+$ T cells (33). Given the experimental support for the hypothesis that IL-7 is critical for the homeostatic proliferation and the survival of naive T cells, in this first model we have neglected other signals (7).

Previous mathematical models of naive T cell homeostasis have focused on the relative contribution of thymic export and cell division in the periphery (3, 34, 35). These models conclude that in humans, thymic export makes an important contribution to the size of the naive T cell population in early life (<20 years of age), whereas later in life the number of naive T cells is maintained by homeostatic proliferation in the periphery. Some recent studies have measured the relative contribution of thymic export by examining the average number of TRECs in naive T cells (3, 35, 36). The use of mathematical models has allowed these groups to infer the relative (young versus old) kinetics for naive T cells, based on experimental estimates of the total number of naive T cells and recent thymic emigrants (3, 34, 35). In the model introduced here, we have aimed to provide a mechanistic perspective by investigating at the molecular and cellular levels, the role IL-7 plays in regulating the homeostasis of naive T cells (37). The increase in the proportion of cycling cells in our model is in agreement with previous experimental studies (38). Our study suggests the increase in peripheral division rates can be explained by the availability of IL-7, which is a consequence of a combined effect: (i) an increased net IL-7 production as an individual ages, and (ii) a reduction of recently exported thymocytes competing for this resource.

ACKNOWLEDGMENTS

We thank Robin Callard, Megan Palmer, and Douglas Lauffenburger for sharing their advice on the approaches taken in this study. This work has been partially supported by the research grant BB/G023395/1 (BBSRC, Carmen Molina-París).

REFERENCES

1. Almeida A, Rocha B, Freitas A, Tanchot C. Homeostasis of T cell numbers: from thymus production to peripheral compartmentalization and the indexation of regulatory T cells. *Semin Immunol* (2005) **17**(3):239–49. doi:10.1016/j.smim.2005.02.002
2. Surh CD, Sprent J. Homeostasis of naive and memory T cells. *Immunity* (2008) **29**(6):848–62. doi:10.1016/j.immuni.2008.11.002
3. Bains I, Thiébaut R, Yates A, Callard R. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**(7):4329–36. doi:10.4049/jimmunol.0900743
4. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, Bregje de Boer A, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* (2012) **36**(2):288–97. doi:10.1016/j.immuni.2012.02.006
5. Takada K, Jameson S. Naive T cell homeostasis: from awareness of space to a sense of place. *Nat Rev Immunol* (2009) **9**(12):823–32. doi:10.1038/nri2657
6. Sprent J, Surh C. Normal T cell homeostasis: the conversion of naive cells into memory-phenotype cells. *Nat Immunol* (2011) **12**(6):478–84. doi:10.1038/ni.2018
7. Tan J, Dudl E, LeRoy E, Murray R, Sprent J, Weinberg K, et al. IL-7 is critical for homeostatic proliferation and survival of naive T cells. *Proc Natl Acad Sci U S A* (2001) **98**(15):8732–7. doi:10.1073/pnas.161126098
8. Schluns K, Kieper W, Jameson S, Lefrançois L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells in vivo. *Nat Immunol* (2000) **1**(5):426–32. doi:10.1038/80868
9. Fry T, Mackall C. The many faces of IL-7: from lymphopoiesis to peripheral T cell maintenance. *J Immunol* (2005) **174**(11):6571–6.
10. Martin B, Bécourt C, Bienvenu B, Lucas B. Self-recognition is crucial for maintaining the peripheral $CD4^+$ T-cell pool in a nonlymphopenic environment. *Blood* (2006) **108**(1):270–7. doi:10.1182/blood-2006-01-0017
11. Brocker T. Survival of mature CD4 T lymphocytes is dependent on major histocompatibility complex class II-expressing dendritic cells. *J Exp Med* (1997) **186**(8):1223–32. doi:10.1084/jem.186.8.1223
12. Markiewicz MA, Brown I, Gajewski TF. Death of peripheral $CD8^+$ T cells in the absence of MHC class I is Fas-dependent and not blocked by Bcl-xL. *Eur J Immunol* (2003) **33**(10):2917–26. doi:10.1002/eji.200324273
13. Vivien L, Benoist C, Mathis D. T lymphocytes need IL-7 but not IL-4 or IL-6 to survive in vivo. *Int Immunol* (2001) **13**(6):763–8. doi:10.1093/intimm/13.6.763
14. Kondrack R, Harbertson J, Tan J, McBreen M, Surh C, Bradley L. Interleukin 7 regulates the survival and generation of memory CD4 cells. *J Exp Med* (2003) **198**(12):1797–806. doi:10.1084/jem.20030735
15. Fry T, Mackall C. Interleukin-7: master regulator of peripheral T-cell homeostasis? *Trends Immunol* (2001) **22**(10):564–71. doi:10.1016/S1471-4906(01)02028-2
16. Mueller SN, Germain RN. Stromal cell contributions to the homeostasis and functionality of the immune system. *Nat Rev Immunol* (2009) **9**(9):618–29. doi:10.1038/nri2588
17. Kang J, Coles M. IL-7: the global builder of the innate lymphoid network and beyond, one niche at a time. *Semin Immunol* (2012) **24**(3):190–7. doi:10.1016/j.smim.2012.02.003
18. Koenen P, Heinzel S, Carrington EM, Happo L, Alexander WS, Zhang J-G, et al. Mutually exclusive regulation of T cell survival by IL-7R and antigen receptor-induced signals. *Nat Commun* (2013) **4**:1735. doi:10.1038/ncomms2719
19. Palmer M, Mahajan V, Chen J, Irvine D, Lauffenburger D. Signaling thresholds govern heterogeneity in IL-7-receptor-mediated responses of naïve $CD8^+$; T cells. *Immunol Cell Biol* (2011) **89**(5):581–94. doi:10.1038/icb.2011.5
20. Stirk ER, Lythe G, van den Berg HA, Molina-París C. Stochastic competitive exclusion in the maintenance of the naïve T cell repertoire. *J Theor Biol* (2010) **265**(3):396–410. doi:10.1016/j.jtbi.2010.05.004
21. Stirk ER, Lythe G, van den Berg HA, Hurst GA, Molina-París C. The limiting conditional probability distribution in a stochastic model of T cell repertoire maintenance. *Math Biosci* (2010) **224**(2):74–86. doi:10.1016/j.mbs.2009.12.004
22. Wallace EW. A simplified derivation of the linear noise approximation. *arXiv Preprint arXiv:1004.4280* (2010).
23. Mitchell WA, Meng I, Nicholson SA, Aspinall R. Thymic output, ageing and zinc. *Biogerontology* (2006) **7**(5–6):461–70. doi:10.1007/s10522-006-9061-7
24. Steinmann GG. Changes in the human thymus during aging. In: Müller-Hermelink HK, editor. *The Human Thymus*. Berlin: Springer (1986). p. 43–88.
25. Haugen F, Norheim F, Lian H, Wensaas AJ, Dueland S, Berg O, et al. IL-7 is expressed and secreted by human skeletal muscle cells. *Am J Physiol Cell Physiol* (2010) **298**(4):C807–16. doi:10.1152/ajpcell.00094.2009
26. Burmaster DE, Crouch EA. Lognormal distributions for body weight as a function of age for males and females in the united states, 1976–1980. *Risk Anal* (1997) **17**(4):499–505. doi:10.1111/j.1539-6924.1997.tb00890.x
27. Geginat J, Lanzavecchia A, Sallusto F. Proliferation and differentiation potential of human CD8+ memory T-cell subsets in response to antigen or homeostatic cytokines. *Blood* (2003) **101**(11):4260–6. doi:10.1182/blood-2002-11-3577
28. Goldrath AW, Bogatzki LY, Bevan MJ. Naive T cells transiently acquire a memory-like phenotype during homeostasis-driven proliferation. *J Exp Med* (2000) **192**(4):557–64. doi:10.1084/jem.192.4.557
29. Saule P, Trauet J, Dutriez V, Lekeux V, Dessaint J-P, Labalette M. Accumulation of memory T cells from childhood to old age: central and effector memory cells in CD4(+) versus effector memory and terminally differentiated memory cells in CD8(+) compartment. *Mech Ageing Dev* (2006) **127**(3):274–81. doi:10.1016/j.mad.2005.11.001

30. Seddon B, Tomlinson P, Zamoyska R. Interleukin 7 and T cell receptor signals regulate homeostasis of CD4 memory cells. *Nat Immunol* (2003) **4**(7):680–6. doi:10.1038/ni946
31. Cho B, Rao V, Ge Q, Eisen H, Chen J. Homeostasis-stimulated proliferation drives naive T cells to differentiate directly into memory T cells. *J Exp Med* (2000) **192**(4):549. doi:10.1084/jem.192.4.549
32. Fry T, Mackall C. Interleukin-7: from bench to clinic. *Blood* (2002) **99**(11):3892–904. doi:10.1182/blood.V99.11.3892
33. Tan JT, Ernst B, Kieper WC, LeRoy E, Sprent J, Surh CD. Interleukin (IL)-15 and IL-7 jointly regulate homeostatic proliferation of memory phenotype CD8+ cells but are not required for memory phenotype CD4+ cells. *J Exp Med* (2002) **195**(12):1523–32. doi:10.1084/jem.20020066
34. Thomas-Vaslin V, Altes HK, de Boer RJ, Klatzmann D. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *J Immunol* (2008) **180**(4):2240–50.
35. Murray JM, Kaufmann GR, Hodgkin PD, Lewin SR, Kelleher AD, Davenport MP, et al. Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty. *Immunol Cell Biol* (2003) **81**(6):487–95. doi:10.1046/j.1440-1711.2003.01191.x
36. Ribeiro RM, Perelson AS. Determining thymic output quantitatively: using models to interpret experimental T-cell receptor excision circle (TREC) data. *Immunol Rev* (2007) **216**(1):21–34. doi:10.1111/j.1600-065X.2006.00493.x
37. Fallon EM, Lauffenburger DA. Computational model for effects of ligand/receptor binding properties on interleukin-2 trafficking dynamics and T cell proliferation response. *Biotechnol Prog* (2000) **16**(5):905–16. doi:10.1021/bp000097t
38. Prelog M, Keller M, Geiger R, Brandstätter A, Würzner R, Schweigmann U, et al. Thymectomy in early childhood: significant alterations of the CD4(+)CD45RA(+)CD62L(+) T cell compartment in later life. *Clin Immunol* (2009) **130**(2):123–32. doi:10.1016/j.clim.2008.08.023
39. Park J, Yu Q, Erman B, Appelbaum J, Montoya-Durango D, Grimes H, et al. Suppression of IL7Rα transcription by IL-7 and other prosurvival cytokines: a novel mechanism for maximizing IL-7-dependent T cell survival. *Immunity* (2004) **21**(2):289–302. doi:10.1016/j.immuni.2004.07.016
40. Henriques CM, Rino J, Nibbs RJ, Graham GJ, Barata JT. IL-7 induces rapid clathrin-mediated internalization and JAK3-dependent degradation of IL-7Rα in T cells. *Blood* (2010) **115**(16):3269–77. doi:10.1182/blood-2009-10-246876
41. Palmer MJ, Mahajan VS, Trajman LC, Irvine DJ, Lauffenburger DA, Chen J, et al. Interleukin-7 receptor signaling network: an integrated systems perspective. *Cell Mol Immunol* (2008) **5**(2):79–89. doi:10.1038/cmi.2008.10

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; accepted: 22 November 2013; published online: 23 December 2013.
*Citation: Reynolds J, Coles M, Lythe G and Molina-París C (2013) Mathematical model of naive T cell division and survival IL-7 thresholds. Front. Immunol. **4**:434. doi: 10.3389/fimmu.2013.00434*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Reynolds, Coles, Lythe and Molina-París. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

A.1. A MODEL OF IL-7R DYNAMICS

In this section we present a summary of a stochastic model in which we consider the number of IL-7 receptors on the surface of naive T cells. This study is used to derive equations (1) and (9) in the main text. The model has been formulated as a continuous time Markov process. We consider the number of receptors on a single naive T cell. For our purposes, we present the mean field approximation to the model with which we estimate the parameters. We introduce the following four variables:

- $m_1(t)$ – the total number of IL-7 receptors on the surface of a naive T cell,
- $m_2(t)$ – the number of unbound internalized receptors,
- $m_3(t)$ – the number of bound internalized receptors,
- $m_4(t)$ – the quantity of IL-7 induced signal.

We assume, in equilibrium, the fraction of surface receptors bound to IL-7 is given by f_I. The system of ODEs denoting the mean field approximation to the stochastic model is given by:

$$\frac{\mathrm{d}m_1(t)}{\mathrm{d}t} = \phi\, \mathrm{e}^{-m_4(t)/\kappa} + \xi_U\, m_2(t) + \xi_B\, m_3(t) - \left[\sigma_U\left(1-f_I\right) + \sigma_B f_I\right] m_1(t)\ , \tag{A1}$$

$$\frac{\mathrm{d}m_2(t)}{\mathrm{d}t} = \sigma_U\left(1-f_I\right) m_1(t) - \left(\xi_U+\zeta_U\right) m_2(t)\ , \tag{A2}$$

$$\frac{\mathrm{d}m_3(t)}{\mathrm{d}t} = \sigma_B f_I\, m_1(t) - \left(\xi_B+\zeta_B\right) m_3(t)\ , \tag{A3}$$

$$\frac{\mathrm{d}m_4(t)}{\mathrm{d}t} = \varphi\, m_3(t) - \chi\, m_4(t)\ . \tag{A4}$$

A.2. REDUCED MODEL IN THE CASE WHEN $I = 0$

Consider a T cell in an IL-7 free medium with initial conditions such that the IL-7 induced signaling vanishes and the number of IL-7:IL-7R internal complexes is zero. Then the mean field model can be reduced to the following set of ODEs:

$$\frac{\mathrm{d}m_1(t)}{\mathrm{d}t} = \phi + \xi_U\, m_2(t) - \sigma_U\, m_1(t)\ ,$$
$$\frac{\mathrm{d}m_2(t)}{\mathrm{d}t} = \sigma_U\, m_1(t) - \left(\xi_U+\zeta_U\right) m_2(t)\,.$$

This reduced system is governed by four parameters ϕ, ξ_U, σ_U, and ζ_U, and possesses the following stable steady state:

$$m_1^* = \frac{\phi\left(\xi_U+\zeta_U\right)}{\sigma_U\zeta_U}\ ,$$
$$m_2^* = \frac{\phi}{\zeta_U}\ .$$

We assume in the reduced model 10% of the total number of receptors are internalized in equilibrium. We set $m_1^* = 9m_2^*$. Based on the measurements of Singer et al. (39), we shall assume 4×10^4 receptors in total when the reduced model is in steady state. Therefore, we let

$$\frac{\phi\left(\xi_U+\zeta_U\right)}{\sigma_U\zeta_U} = 3.6\times10^4, \tag{A5}$$

$$\frac{\phi}{\zeta_U} = 4\times10^3\,. \tag{A6}$$

In Ref. (40), cells were cultured with the translation inhibitor cycloheximide (CHX) to prevent transcription of the IL-7 receptor. Total expression of the IL-7 receptor was measured over several time points, from which the authors estimate the half-life of the receptor in an unstimulated cell to be approximately 24 h.

In the reduced model, all receptors are guaranteed to be degraded in a finite amount of time. The expected time for a receptor, that is initially on the cell surface, to be degraded in the lysosome is given by

$$\tau_1 = \frac{\zeta_U+\sigma_U+\xi_U}{\sigma_U\zeta_U}\ .$$

Assuming exponential decay, the half-life for a receptor to undergo lysosomal degradation, starting on the cell surface, is then given by

$$t_{\frac{1}{2}} = \frac{\zeta_U+\sigma_U+\xi_U}{\sigma_U\zeta_U}\log 2\,.$$

Thus, we can write

$$\frac{\zeta_U+\sigma_U+\xi_U}{\sigma_U\zeta_U}\log 2 = 24\ \mathrm{h}\,. \tag{A7}$$

Combining equations (A5), (A6), and (A7) we find

$$\zeta_U \approx 0.29\ \mathrm{h}^{-1}\,,$$
$$\phi \approx 1.2\times10^3\ \text{receptors}\ \mathrm{h}^{-1}\,,$$
$$\xi_U + 0.29\ \mathrm{h}^{-1} \approx 9\sigma_U.$$

The value of ξ_U relative to ζ_U effectively dictates the ratio of receptors which are degraded to those recycled back to the cell surface. We assume the system has evolved to minimize waste of functional proteins and tentatively let $\xi_U > \zeta_U$. That is, we assume that a greater fraction of receptors are recycled back to the surface of the cell. We somewhat arbitrarily set

$$\xi_U = 1\ \mathrm{h}^{-1} \quad\Rightarrow\quad \sigma_U \approx 0.14\ \mathrm{h}^{-1}\,.$$

A.3. RECEPTOR-LIGAND KINETICS

Suppose the number of surface receptors is constant and denoted by R_T. Let us also assume the extra-cellular concentration of IL-7 is constant and denoted by I. Define $R_B(t)$ to be the number of IL-7 receptors bound to IL-7. Note that we assume the time to recruit the common gamma chain, γ_c, is negligible. Then, we can describe changes in the number of bound complexes by the following ODE

$$\frac{\mathrm{d}R_B(t)}{\mathrm{d}t} = k_+\left[R_T - R_B(t)\right] I - k_- R_B(t)\ ,$$

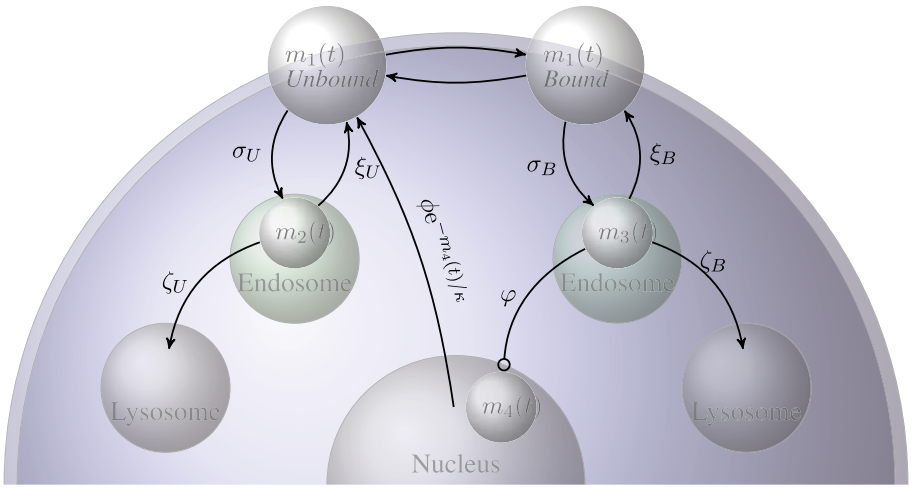


FIGURE A1 | Diagrammatic representation of the transition probabilities of the stochastic model for the IL-7 and IL-7 receptor system.

where k_+ and k_- are, respectively, the binding and unbinding rates of the IL-7:IL-7R receptor-ligand system. We assume the timescales for this reaction are faster than the timescales for changes in receptor numbers, such that we can consider these reactions to be in equilibrium. This ODE has a unique stable steady state:

$$R_B^* = \frac{R_T I}{\frac{k_-}{k_+} + I} = f_I R_T \, .$$

The Appendix for Ref. (41) provides estimates for k_+ and k_-, from which we set $k_+ = 1\,\text{nM}\,\text{min}^{-1}$ and $k_- = 0.1\,\text{min}^{-1}$. It is reported IL-7 has a molecular mass of around 17 kDa, from which we estimate the ratio $k_-/k_+ \approx 1.7\,\text{ng}\,\text{ml}^{-1}$.

A.4. EARLY INTERNALIZATION EVENTS

In Ref. (40), surface receptor expression was assessed in human thymocytes. It is reported, cells in $50\,\text{ng}\,\text{ml}^{-1}$ IL-7 culture down-regulated IL-7R expression. A 20% reduction was observed after 10 min. Using the above estimate for k_-/k_+, we find $f_I|_{I=50} \approx 0.97$. Based on this, we can neglect internalization of the unbound receptor. In the first 10 min, we shall also neglect recycling and inhibition of receptor transcription. We assume surface receptor expression loss is modeled by the ODE

$$\frac{\mathrm{d}m_1(t)}{\mathrm{d}t} = \phi - \sigma_B m_1(t) \, .$$

Given initial surface expression levels equal to $m_1(0)$, this ODE has solution

$$m_1(t) = \frac{\phi}{\sigma_B} + \left[m_1(0) - \frac{\phi}{\sigma_B} \right] \mathrm{e}^{\sigma_B t} \, . \tag{A8}$$

We assume the previous estimates obtained from the reduced model for $m_1(0)$ and ϕ. That is, we let $m_1(0) = 3.6 \times 10^4$ receptors and $\phi = 1.2 \times 10^3$ receptors h^{-1}. Then using the above expression for $m_1(t)$ equation (A8), we obtain an estimate for σ_B. We find $\sigma_B \approx 1.4\,\text{h}^{-1}$. The authors of Ref. (40) estimate the half-life of the IL-7 receptor in cells treated with CHX, cultured in $50\,\text{ng}\,\text{ml}^{-1}$, to be approximately 3 h. The expected time to lysosomal degradation for a surface receptor is given by

$$\begin{aligned}\tau_2 &= \frac{\left[\sigma_U\left(1-f_I\right) + \left(\xi_U + \zeta_U\right)\right]\left(\xi_B + \zeta_B\right) + \sigma_B f_I \left(\xi_U + \zeta_U\right)}{\sigma_U\left(1-f_I\right)\zeta_U\left(\xi_B + \zeta_B\right) + \sigma_B f_I \left(\xi_U + \zeta_U\right)\zeta_B} \\ &\approx \frac{1.3\,\text{h}^{-1}\left(\xi_B + \zeta_B\right) + 1.7\,\text{h}^{-2}}{4.6 \times 10^3\,\text{h}^{-2}\left(\xi_B + \zeta_B\right) + 1.7\,\text{h}^{-2}\zeta_B} .\end{aligned}$$

Assuming exponential decay, with a half-life of 3 h, we find $\xi_B \approx 0.2\zeta_B + 0.3\,\text{h}^{-1}$. Again, without a direct measurement, let us set $\xi_B = 1\,\text{h}^{-1}$, to obtain an estimate for $\zeta_B \approx 3.5\,\text{h}^{-1}$.

A.5. REMAINING PARAMETERS

Consider the observation of an approximately 98% reduction in surface receptor expression following overnight culture in $6\,\text{ng}\,\text{ml}^{-1}$ IL-7 made in Ref. (39). We let $m_1^*|_{I=6} = 0.02\ m_1^*|_{I=0} = 0.02\frac{\phi(\xi_U+\zeta_U)}{\sigma_U\zeta_U} \approx 763$ receptors. We set the derivatives equal to zero in the system of ODEs equations (A1–A4), and manipulate the resulting system of equations to obtain

$$\begin{aligned}&\phi \exp\left(-\frac{\varphi\sigma_B f_I}{\chi\kappa\left(\xi_B + \zeta_B\right)} m_1^*\right) \\ &\quad = \left[\sigma_U\left(1-f_I\right) + \sigma_B f_I - \frac{\sigma_U\left(1-f_I\right)\xi_U}{\xi_U + \zeta_U} - \frac{\sigma_B f_I \xi_B}{\xi_B + \zeta_B}\right] m_1^* \, , \end{aligned} \tag{A9}$$

$$\Rightarrow \quad \exp\left(-1.9 \times 10^2\ \text{receptors}\ \frac{\varphi}{\chi\kappa}\right) \approx 0.54 \, ,$$

$$\Rightarrow \quad \frac{\varphi}{\chi\kappa} \approx 3.2 \times 10^{-3}\ \text{receptors}^{-1}.$$

Table A1 | Parameter estimates obtained from the mean field model of IL-7 receptor dynamics.

Parameter	Value	Units
ϕ	1.2×10^3	rec h^{-1}
κ	10^3	sig
ξ_U	1	h^{-1}
ξ_B	1	h^{-1}
σ_U	0.14	h^{-1}
σ_B	1.4	h^{-1}
k_-/k_+	1.7	ng ml^{-1}
ζ_U	0.29	h^{-1}
ζ_B	3.5	h^{-1}
φ	0.61	sig rec^{-1} h^{-1}
χ	0.19	h^{-1}

We let sig be the units of the signaling and rec be the units of surface receptor numbers.

From the steady solutions of ODEs equations (A3) and (A4), we find

$$m_4^* = \frac{\phi\sigma_B f_I}{(\xi_B + \zeta_B)\,\chi} m_1^* \approx 1.9 \times 10^2 \text{ receptors } \frac{\varphi}{\chi}.$$

We use this expression to rearrange equation (A9) in terms of m_4^*. This gives

$$\exp\left(-\frac{m_4^*}{\kappa}\right) \approx 9.1\frac{m_4^*}{\kappa}\,.$$

Solving the above expression numerically, we find $m_4^*/\kappa \approx 0.1$. We estimate a value for χ based on the observation that following culture in 6 ng ml^{-1} IL-7, mRNA levels took approximately 12 h to return to 99% of the control levels. The transcription rate is given by $\phi\exp(-m_4(t)/\kappa)$. We again assume the IL-7 induced signal decays according to the equation $m_4(t) = m_4(0)e^{-\chi t}$, where $m_4(0) = 0.1$. Combing these assumptions with the experimental observations, we have

$$\phi \exp\left[-\frac{m_4\,(0)\exp\,(-12\chi)}{\kappa}\right] = 0.99\phi,$$

from which we find $\chi \approx 0.19\,h^{-1} \Rightarrow \varphi \approx 6.1 \times 10^{-4}$ receptors^{-1} h^{-1} κ. The parameter κ was chosen to be 1000. This choice was made from the stochastic model: $\kappa = 1000$ is the minimum value (to the nearest power of 10) such that fluctuations in the signaling quantity are greater than zero for low (10^{-2} ng ml^{-1}) concentrations of IL-7. Using this value for κ, we find $\phi \approx 0.61\,h^{-1}$. The parameter estimates are summarized in **Table A1**.

A.5.1. Changes in the concentration of IL-7

The functional form

$$S = \frac{aI}{b + I} \tag{A10}$$

is used to approximate the numerical solution of m_4^* as a function of the concentration of IL-7. Using this function we find $a \approx 600$ signaling units and $b \approx 0.025$ ng ml^{-1}. In a similar manner we use the functional form

$$R = c + \frac{d}{e + I} \tag{A11}$$

to approximate the numerical solution of m_1^* as a function of the concentration of IL-7. We find $c \approx 600$ receptors, $d \approx 1000$ receptors ng ml^{-1} and $e \approx 0.03$ ng ml^{-1}. The number of IL-7 molecules internalized by each T cell per day is assumed to be the same as the number of internalized IL-7:IL-7R complexes per day. We let the complex internalization rate, as a function of I, be given by

$$24\sigma_B f_I\, m_1^*\,(I) \approx \frac{6.7I}{2 + I}\left(3 + \frac{5}{3 \times 10^{-2} + I}\right) \times 10^3 \text{ molecules cell}^{-1}\text{ day}^{-1}. \tag{A12}$$

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 December 2013
doi: 10.3389/fimmu.2013.00474

A mathematical model of immune activation with a unified self-nonself concept

Sahamoddin Khailaie[1], Fariba Bahrami[2], Mahyar Janahmadi[3], Pedro Milanez-Almeida[4], Jochen Huehn[4] and Michael Meyer-Hermann[1,5]*

[1] *Department of Systems Immunology, Helmholtz Centre for Infection Research, Braunschweig, Germany*
[2] *CIPCE, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[3] *Neuroscience Research Centre and Department of Physiology, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[4] *Department of Experimental Immunology, Helmholtz Centre for Infection Research, Braunschweig, Germany*
[5] *Bio Centre for Life Science, Technische Universität Braunschweig, Braunschweig, Germany*

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Ichiro Taniuchi, RIKEN Research Center for Allergy and Immunology, Japan
Robin Callard, University College London, UK
Grant Lythe, University of Leeds, UK

***Correspondence:**
Michael Meyer-Hermann, Department of Systems Immunology, Helmholtz Centre for Infection Research, Inhoffenstr.7, 38124 Braunschweig, Germany
e-mail: mmh@theoretical-biology.de

The adaptive immune system reacts against pathogenic nonself, whereas it normally remains tolerant to self. The initiation of an immune response requires a critical antigen(Ag)-stimulation and a critical number of Ag-specific T cells. Autoreactive T cells are not completely deleted by thymic selection and partially present in the periphery of healthy individuals that respond in certain physiological conditions. A number of experimental and theoretical models are based on the concept that structural differences discriminate self from nonself. In this article, we establish a mathematical model for immune activation in which self and nonself are not distinguished. The model considers the dynamic interplay of conventional T cells, regulatory T cells (Tregs), and IL-2 molecules and shows that the renewal rate ratio of resting Tregs to naïve T cells as well as the proliferation rate of activated T cells determine the probability of immune stimulation. The actual initiation of an immune response, however, relies on the absolute renewal rate of naïve T cells. This result suggests that thymic selection reduces the probability of autoimmunity by increasing the Ag-stimulation threshold of self reaction which is established by selection of a low number of low-avidity autoreactive T cells balanced with a proper number of Tregs. The stability analysis of the ordinary differential equation model reveals three different possible immune reactions depending on critical levels of Ag-stimulation: a subcritical stimulation, a threshold stimulation inducing a proper immune response, and an overcritical stimulation leading to chronic co-existence of Ag and immune activity. The model exhibits oscillatory solutions in the case of persistent but moderate Ag-stimulation, while the system returns to the homeostatic state upon Ag clearance. In this unifying concept, self and nonself appear as a result of shifted Ag-stimulation thresholds which delineate these three regimes of immune activation.

Keywords: immune activation, autoimmunity, autoreactive T cells, regulatory T cells, central tolerance, peripheral tolerance

INTRODUCTION

The immune system is continuously exposed to a wide variety of disturbances. Such disturbances are recognized by T cells via antigen presentation. Antigen presentation is a process in which antigen presenting cells (APC) capture the antigens, break them into small peptides, couple them with MHC molecules, and present them on the cell surface, thus enabling their recognition by T cells (1–3). The majority of disturbances that the immune system deals with are pathogenic nonself-antigens. Since the APCs break down the nonself-antigens into smaller peptides and present them on their surface, presented peptide of nonself might have overlaps with self-peptides (4, 5).

In addition, rapidly evolving nonself pathogens, such as Hepatitis C virus, might acquire similarities to self-antigens (6). Apart from nonself, altered self such as cancer cells is also a disturbance that has to be recognized by the immune system. Therefore, an ideal immune system has to find a solution for dealing with nonself, self-similar nonself, and self-disturbances (**Figure 1**).

As a general solution, the immune system generates T cell clones with random specificities that could potentially recognize any peptides, including self-peptides. The classical idea that the T cell repertoire has to be self-tolerant and T cells should not react to self-peptides, assumes that self-reactive T cells should be eliminated. This assumption is partially true, as T cell clones which fully recognize self-peptides in the thymus undergo clonal deletion, in the so-called negative selection process (7, 8).

The self-tolerance resulting from negative selection is called central tolerance. A stringent central tolerance induction and deletion of all autoreactive T cells is believed to create holes in the specificity space of the T cell repertoire (9, 10) by prohibiting immune responses against self-similar nonself and altered self. Hence, a too stringent central tolerance does not seem beneficial.

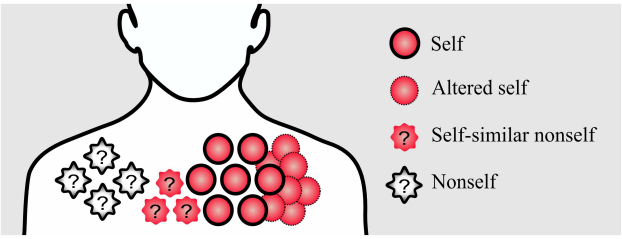


FIGURE 1 | Conceptual figure of different disturbances in the immune system. Pathogenic nonself-disturbances are recognized and attenuated by nonself-specific T cells. However, recognition and attenuation of altered self and self-similar nonself-disturbances is challenging for the immune system due to the existence of self-tolerance mechanisms; without self-specific immune cells, the immune system is not able to initiate an immune response against these disturbances.

In line with this idea, there is evidence that negative selection only partially deletes autoreactive T cells because availability of self-peptides required for negative selection in the thymus is limited and T cells spend only a limited time in the process of thymic selection (11–13). Autoreactive T cells escaping negative selection have been shown to be involved in autoimmunity (14). They normally exist in healthy individuals and are quiescent in steady state conditions in the presence of their cognate self-antigen (15).

Escaped autoreactive T cells are under the control of peripheral tolerance. A prominent mechanism of peripheral tolerance among others [reviewed in Ref. (16)] is induced by $CD4^+$ $Foxp3^+$ regulatory T cells (Tregs) (17). The majority of these cells, known as natural Tregs, are hypothesized to be selected from autoreactive T cells in thymus (18, 19). The main role of Tregs is the regulation of the immune response by suppression of the effector functions of conventional T cells (Tconv).

Despite the necessity of suppression by Tregs for avoiding autoimmunity (20, 21), production of too large numbers of Tregs in the thymus might prevent beneficial effector responses. Therefore, a too stringent peripheral tolerance induction by selection of large numbers of Tregs in the thymus does not seem favorable.

In view of this background, how does the immune system balance the tolerance mechanisms in order to ensure immune responses to any kind of disturbances including self-disturbances, yet staying tolerant to self in healthy homeostasis? Here, we address this question by using a mathematical model of immune activation that relies on identical components for self and nonself, i.e., using the same set of ordinary differential equations. The model considers the thymic production of Tregs and Tconvs as well as the dynamic interplay between Tregs, Tconvs, and IL-2 molecules in the presence of antigen(Ag)-stimulation in the periphery. The model is exploited to reveal the parametric regime of the immune system in which an immune response against self is restricted, but not impossible.

The interplay between Tregs and Tconvs during immune responses is a topic of extensive mathematical modeling (22–28). León and co-workers (22) proposed a series of models for studying immune tolerance by considering APCs, Tconvs, and Tregs. Their models rely on the assumption that regulatory interaction between Tregs and Tconvs takes place only in simultaneous conjugation with an APC. As a result of this assumption, efficient suppression of Tconvs requires a minimum population of Tregs per APC (29). As an extension, a crossregulation model is proposed by Carneiro and co-workers (26) in an attempt to incorporate Tregs in a coherent theory of the immune system. According to their model that shows a bistable behavior, immunity to a given Ag arises as competitive exclusion of Tregs by the expansion of Tconvs and tolerance results from limited APC availability or above threshold Treg numbers. Since the interactions between Tregs and Tconvs is assumed to depend on the density of the APCs, increasing the APC availability decreases the simultaneous conjugate formation of Tregs and Tconvs with the same APCs and hence, it is sufficient to break the immune tolerance.

An alternative concept was brought forward in a model proposed by Carneiro and co-workers (23) that assumes APC-independent interactions between Tconvs and Tregs for immune suppression which will be also used in our model. A direct interaction of Tconvs and Tregs was identified by experiments (30). The authors concluded that efficient immune suppression still requires a minimum population of Tregs regardless of the number of APCs.

In contrast to the aforementioned studies, we do not consider the conjugate formation of Tregs and Tconvs with APCs and therefore, there is not a competition between these cells for Ag. Instead, the role of APCs is indirectly captured by an Ag-stimulation factor which is the activation rate of naïve T cells and resting Tregs with identical Ag-specificity by APCs bearing their cognate Ag. In addition, we explicitly consider the dependency of Tregs on Tconvs through the growth factor IL-2.

Burroughs and co-workers (24) investigated Treg-induced inhibition of cytokine secretion by effector T cells. By assuming that Tregs are activated by self Ag and locally maintained by nonlinear competition for tissue-derived cytokines that are solely utilized by Tregs, the authors analyzed the role of local active Treg population size in the balance between suppressor and effector responses. Stimulation of Tregs and Tconvs is described by independent parameters. In contrast to their model, thymic output maintains the homeostatic population of Tregs in our model. Another essential difference is that Ag-stimulation of Tregs and Tconvs is described with a unified self-nonself concept and Tregs are assumed to also respond to nonself Ag-stimulation (31).

Parametric steady state analysis of the model provides insights about the physiological range of model parameters, and determines the overall conditions under which immune responses against self are possible. Furthermore, the impact of model parameters on the requirements for the initiation of immune reactions against self is analyzed. The model proposes that disturbed homeostatic balance between autoreactive T cells and Tregs increases the susceptibility to autoimmunity or cancer.

RESULTS

The mathematical model is constructed starting from a simple model of the immune response including essential components only. Then, additional complexity is incrementally added to the model to a degree allowing for validation and analysis of tolerance versus immunity. The scheme of the complete model is

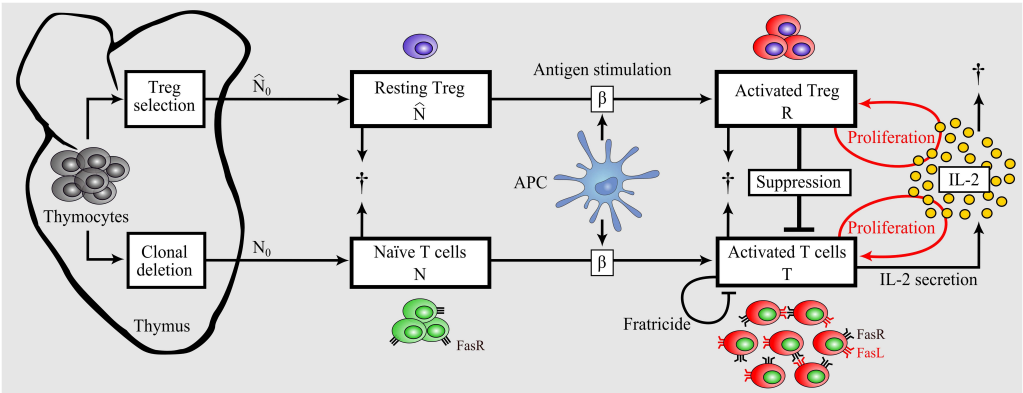


FIGURE 2 | Model of dynamic interplay between conventional T cells and regulatory T cells. Nonself-specific as well as some self-specific thymocytes that survived negative selection and were not selected as Tregs enter the periphery as naïve T cells. A part of detected autoreactive thymocytes differentiate into Tregs in the thymus and reside in the periphery in resting state. Upon Ag-stimulation by APCs, naïve T cells and resting Tregs become activated. In contrast to activated T cells, activated Tregs do not secrete IL-2, but both activated populations proliferate in dependence on the presence of IL-2 (46). Activated Tregs suppress activated T cells in a cell-contact-dependent and cytokine-driven manner. Activated T cells undergo Fas-induced apoptosis by interacting with each other (fratricide). In contrast, Tregs are resistant to Fas-induced apoptosis (68).

depicted in **Figure 2**. The model is conceptually independent of the self/nonself nature of the immune response, and differences of the immune responses against self versus nonself are reflected in different parameter values of the same model.

AN IMMUNE RESPONSE REQUIRES SUFFICIENT DIVISION AND IL-2 SECRETION RATE OF ACTIVATED T CELLS

Immune responses arise from massive proliferation of activated T cells and their subsequent effector function. Our simplest model attempts to capture the dynamic characteristics of an activated T cell population (T):

$$\begin{cases} \dfrac{\mathrm{d}T}{\mathrm{d}t} = aIT - bT \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = dT - eIT - fI \end{cases} \tag{1}$$

Activated T cells have a mean lifespan $1/b$ and secrete IL-2 (I) with rate d. Available IL-2 decays with rate f and is consumed by activated T cells with rate e. Activated T cells are able to proliferate (with rate aI) in the presence of IL-2. This IL-2 dependent proliferation rate is considered as a linear function of IL-2 in model (1). The impact of considering a nonlinear proliferation rate (a Hill-function of IL-2) instead of the linear term aIT is given in Section "Nonlinear Proliferation Rate of Conventional and Regulatory T Cells" in Appendix.

Steady state analysis of the model (1) is given in Section "Steady State Analysis of Model (1)" in Appendix. This model has two equilibrium points:

$$(T_1, I_1) = (0, 0), (T_2, I_2) = \left(\frac{bf}{ad - be}, \frac{b}{a}\right) \tag{2}$$

By assuming the biological range of parameters (all parameters are positive), the trivial equilibrium point (T_1, I_1) is stable and the nontrivial equilibrium (T_2, I_2) is unstable. T_2 is positive if and only if:

$$ad - be > 0 \tag{3}$$

The unstable equilibrium point imposes a threshold for initial conditions of the model in which the activated T cells proliferate unlimitedly, which in this simplest model, corresponds to an efficient immune response. This can be visualized by the phase portrait of the model as shown in **Figure 3A**. The condition (3) imposes a quality constraint on activated T cell clones to enter a highly proliferative state and implies that among T cell clones that are in the activated state, only the T cell clones with a sufficiently high proliferation rate (a) or IL-2 secretion rate (d) are able to contribute to the immune response against Ag. Since both, the proliferation and IL-2 secretion rate of activated T cells depend on the affinity/avidity of their TCR to the presented Ag (32–34), condition (3) implies that the existence of T cell clones with sufficiently high specificity for the presented Ag is required for induction of an immune response. Similar implications were derived from a model that considers a nonlinear IL-2 dependent proliferation rate of activated T cells (Nonlinear Proliferation Rate of Conventional and Regulatory T Cells in Appendix).

The major focus of central tolerance is to eliminate T cells that are self-specific. Therefore, it is unlikely that highly self-specific T cells escape from central tolerance, as they are more effectively detected and eliminated in the thymus (12, 34). It is thus expected that autoreactive T cells in the periphery are less aggressive than the ones that undergo clonal deletion in the thymus, and may not fulfill condition (3).

INITIATION OF AN IMMUNE RESPONSE REQUIRES A MINIMUM HOMEOSTATIC POPULATION OF NAÏVE T CELLS AND ANTIGEN STIMULATION

Continuous thymic production of naïve T cells maintains the peripheral number and diversity of mature naïve T cells (35),

although other mechanisms such as stimulation of T cells with self-antigens and IL-7 have been shown to be involved (36). Upon Ag-stimulation by activated APCs, naïve T cells with high avidity to the presented Ag become activated. Here, we take into account the dynamics of the naïve T cell population (N) and T cell activation by Ag-stimulation (β), as described in equations (4). We assume that naïve T cells with identical Ag-specificity have a homeostatic population in the periphery that is established by naïve T cell renewal (by rate N_0) and natural cell death (with rate g):

$$\begin{cases} \frac{dN}{dt} = f(N) = N_0 - gN - \beta N \\ \frac{dT}{dt} = aIT - bT + \beta N \\ \frac{dI}{dt} = dT - eIT - fI \end{cases} \quad (4)$$

T cell activation $k(t)$ is defined as

$$k(t) = \beta N(t) \quad (5)$$

Steady state analysis of model (4) is given in Section "Steady State Analysis of Model (4)" in Appendix. This model has either 2 or no equilibrium points dependent on the steady state value of T cell activation (k). According to the bifurcation diagram of the model depicted in **Figure 3B**, which is obtained by treating k as bifurcation parameter, model (4) has no equilibrium points for:

$$k > k_{-} = \frac{adf}{e^2}\left(1 - \sqrt{1 - \frac{be}{ad}}\right)^2 \quad (6)$$

which corresponds to the unlimited proliferation state of activated T cells. Therefore, condition (6) has to be satisfied for initiation of an immune response. However, according to model (4), the steady state value of T cell activation (k) is limited by naïve T cell renewal (N_0) and Ag-stimulation (β):

$$k = \frac{\beta N_0}{g + \beta} \quad (7)$$

Therefore, according to equations (6) and (7), there exists an Ag-stimulation range

$$\beta > \frac{gk_{-}}{N_0 - k_{-}} \quad (8)$$

in which an immune response is initiated if:

$$N_0 > k_{-} \quad (9)$$

Condition (9) implies that the renewal rate of naïve T cells plays a critical role for the initiation of immune responses. In other words, without a sufficient renewal rate of naïve T cells, the immune response cannot be initiated by any Ag-stimulation. Instead, Ag-stimulation results in a subcritical immune response which is interpreted as insufficient for pathogen clearance. By increasing the proliferation rate or IL-2 secretion of activated T cells or the renewal rate of naïve T cells, the threshold of Ag-stimulation required for initiation of an immune response is decreased [equations (6) and (8)]. Therefore, central tolerance is able to tune the initiation criterion of self reaction not only by limiting the quality of autoreactive T cells, but additionally by restricting the renewal rate of autoreactive T cells. As central tolerance does not limit nonself-specific T cells, according to the model, these cells exhibit a lower threshold of activation by nonself Ag-stimulation.

FRATRICIDE: A MECHANISM TO LIMIT BUT NOT TO SUPPRESS IMMUNE RESPONSES

The immune response in model (4) is characterized by unlimited proliferation of activated T cells which is physiologically unrealistic. The linear death term of natural death of activated T cells in model (4) is not sufficient to limit proliferation, and requires a nonlinear limiting factor. A potential mechanism of limiting the immune response is activation-induced cell death (AICD) in activated T cells, known as fratricide (37). Upon T cell activation, death ligand (FasL) and receptor (Fas) proteins are expressed on

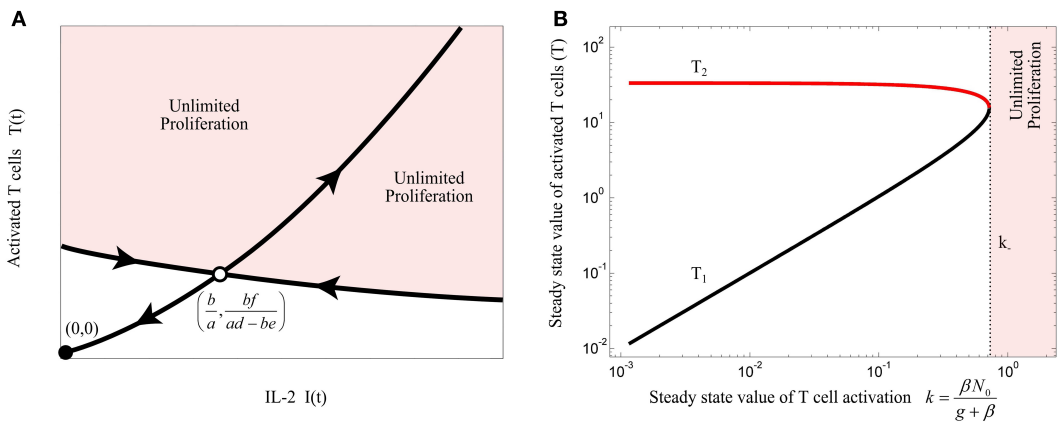


FIGURE 3 | (A) Qualitative phase portrait of model (1): the stable manifold of saddle node defines a threshold for the initial conditions that allow for unlimited proliferation of activated T cells. **(B)** Bifurcation diagram of model (4) by treating k as bifurcation parameter. Stable and unstable equilibrium points are shown by black and red lines, respectively. For $k > k_{-}$, the immune response enters the regime of unlimited proliferation.

the surface of T cells. Followed by expression of these proteins, T cells eliminate themselves in a cell-contact-dependent manner. The fratricide mechanism is modeled by a nonlinear death term (cT^2), as proposed by Callard et al. (37):

$$\begin{cases} \dfrac{dN}{dt} = f(N) = N_0 - gN - \beta N \\ \dfrac{dT}{dt} = aIT - bT - cT^2 + \beta N \\ \dfrac{dI}{dt} = dT - eIT - fI \end{cases} \tag{10}$$

The steady state analysis of model (10) is provided in Section "Steady State Analysis of Model (10)" in Appendix. This model has either 3 or 1 equilibrium points, depending on the value of fratricide death rate c. The bifurcation diagram of the model (10) with respect to c is depicted in **Figure 4A** for ($\beta = 0$). When c satisfies

$$c < c_- = f^{-1}\left(\sqrt{ad} - \sqrt{be}\right)^2 \tag{11}$$

the stable equilibrium point (T_3) exists and corresponds to a saturated population of activated T cell. When the conditions (3) and (11) are fulfilled, the model (10) exhibits the bifurcation diagram plotted in **Figure 4B** with respect to the steady state value of T cell activation (k). The fratricide mechanism added a large stable equilibrium point (T_3) to the model which imposes a saturation level to the activated T cell population. The larger the c, the smaller the saturated population of activated T cells is. Similar to model (4), model (10) shows an initiation threshold of the immune response ($k > k_i$). Despite solving the issue of unlimited proliferation of activated T cells by the fratricide mechanism, model (10) bears a hysteresis characteristic so that the immune response cannot be suppressed when Ag-stimulation (β) is removed.

DYNAMIC INTERPLAY OF ACTIVATED T CELLS AND TREGS

Tregs are essential in maintaining self-tolerance and immune homeostasis by preventing autoimmunity and limiting chronic inflammation in the periphery. However, they might also limit beneficial responses by inducing tolerance to pathogens (38, 39) or limiting anti-tumor immunity (40, 41). One functional role of Tregs is to shut down the cell-mediated immune response via cell-contact-dependent and inhibitory cytokine-driven suppression of activated T cells (42). Two different subsets of Tregs were identified. Natural Tregs are the dominant subset of peripheral Tregs (43) and are selected in the thymus. In our model, we consider only natural Tregs and neglect the induced Treg subset that differentiates from naïve T cells. Like for naïve T cells, the thymus contributes to the renewal of resting Tregs ($\widehat{N}$) by continuously selecting them from thymocytes. The renewal of resting Tregs is assumed to occur by rate $\widehat{N}_0$. Since we are interested in the relative renewal of resting Tregs and naïve T cells, we assume that:

$$\widehat{N}_0 = \lambda N_0 \tag{12}$$

Tregs remain in the resting state until they are stimulated by Ag (β) and become activated in a TCR-dependent manner. The dynamic population of the resting Treg compartment is assumed to be the same as the naïve T cell compartment in (4) and (10) ($d\widehat{N}/dt = f(\widehat{N})$). Activated Tregs ($R$) are assumed to suppress activated T cells (by rate γ). Survival and proliferation of activated Tregs depends strictly on IL-2, produced by activated non-Tregs (44–46). The IL-2 dependent proliferation rate of Tregs is considered as a linear function of IL-2 (see Nonlinear Proliferation Rate of Conventional and Regulatory T Cells in Appendix for a nonlinear case). In contrast to activated T cells, activated Tregs

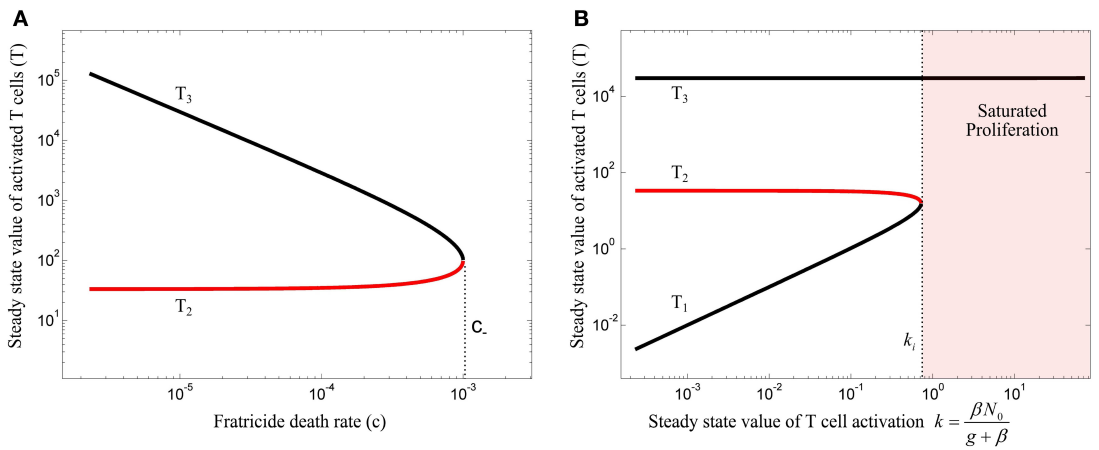


FIGURE 4 | (A) Bifurcation diagram of model (10) with $\beta = 0$ using the fratricide death rate c as bifurcation parameter. No immune response exists for fratricide death rates larger than c_- due to extensive activation-induced cell death. The trivial equilibrium point is omitted in this figure. **(B)** Bifurcation diagram of model (10) using k as bifurcation parameter: an immune response can be initiated for large values of k. However, due to hysteresis characteristic in this model, the immune response is not suppressed after decreasing T cell activation (k). Stable and unstable equilibrium points are shown by black and red lines, respectively.

lack the ability to secrete IL-2 (47). The relative proliferation rate of activated Tregs and activated T cells is controlled by the parameter ϵ:

$$\begin{cases} \dfrac{dT}{dt} = aIT - bT - cT^2 - \gamma RT + \beta N \\ \dfrac{dR}{dt} = \epsilon aIR - bR + \beta \widehat{N} \\ \dfrac{dI}{dt} = dT - eI(T+R) - fI \end{cases} \tag{13}$$

The parameters are given in **Table 1** and the model components are illustrated in **Figure 2**. Treg activation $\widehat{k}(t)$ is defined as

$$\widehat{k}(t) = \beta \widehat{N}(t) \tag{14}$$

According to equations (7) and (12), the steady state value of Treg activation ($\widehat{k}$) is given by

$$\widehat{k} = \lambda \frac{\beta N_0}{g+\beta} = \lambda k \tag{15}$$

The equilibrium points of model (13) are given in Section "Steady State Analysis of Model (13)" in Appendix. By incorporating the Treg compartment to model (10), two additional equilibrium points (T_4 and T_5) emerged for $\beta = 0$. The equilibrium point of interest (T_4), which depends on the Treg-associated parameters (ϵ, γ), has an impact on the topological changes of the phase portraits of the model under variations of the bifurcation parameter k. The value of ϵ and γ are assumed to be in a range where the model does not inherit the hysteresis characteristics of immune responses from model (10) in which the immune response is not suppressed after resolving Ag-stimulation (β). Then, the bifurcation diagrams of model (13) for two different values of λ are obtained by treating k as the bifurcation parameter (**Figure 5**). Depending on the value of k, the model has either 5 or 3 equilibrium points.

By varying the relative renewal rates of resting Tregs and naïve T cells [λ in equation (12)] a λ_{th} can be found, so that no immune response can be initiated for any value of k, if $\lambda > \lambda_{th}$ (**Figure 5A**). For $\lambda < \lambda_{th}$ (**Figure 5B**), there exists a T cell activation threshold (k_i) such that for $k > k_i$ the immune response can be initiated. However, in contrast to model (10), the immune response is completely suppressed by activated Tregs if k decreases to a lower value than k_i (gray region in **Figure 5B**). For persistent Ag-stimulation with $k > k_i$, two scenarios are possible. An oscillating immune response is induced when k remains in the range of $k_i < k < k_s$ (red region in **Figure 5B**). For $k > k_s$ the immune response is suppressed after its initiation to a minor immune response with an activated T cell population T_4 due to over-suppression of activated T cells by over-activation of Tregs (magenta region in **Figure 5B**). In the latter case ($k > k_s$), despite proper T cell stimulation, only a minor immune response is induced (and antigen is not cleared). Instead a chronic co-existence of antigen and inefficient immune activity is established. Therefore, according to the model, a range of T cell and Treg activation ($k_i < k < k_s$) exists in which an efficient immune response is induced. Outside of this range, the antigen persists because of under-stimulation of naïve T cells, or over-stimulation of Tregs. According to equation (7), the existence of Ag-stimulation thresholds β_i and β_s which correspond to the values of k_i and k_s, respectively, depends on the renewal rate of naïve T cells (N_0); β_i exists if $N_0 > k_i$ and β_s exists if $N_0 > k_s$. Increasing the renewal rate of naïve T cells reduces the Ag-stimulation required for initiation(β_i)/over-suppression(β_s) of the immune response.

The peak immune response depends on the value of the Treg-associated equilibrium point (T_4) which in turn depends on Treg-associated parameters. However, the fratricide-associated equilibrium point (T_3) is a limiting factor for the maximum population of activated T cells if the fratricide death rate (c) is sufficiently high.

According to our model, sufficient activated Tregs are required to suppress the proliferative response of activated T cells. These are supplied by two processes: Treg activation ($\widehat{k}$) which depends on Ag-stimulation (β), and Treg proliferation which depends on the IL-2 secretion by activated T cells. With a low Ag-stimulation and insufficient Treg activation ($\widehat{k} = \beta \widehat{N}$), Treg proliferation has to account for immune suppression. Since Treg proliferation is

Table 1 | Parameters used for model analysis.

Parameter	Value	Description	Dimension
a	0.4	Proliferation rate of activated T cells	molecules^{-1}time^{-1}
b	0.1	Natural death rate of activated T cells and Tregs	mime^{-1}
c	10^{-5}	Fratricide death rate of activated T cells	cells^{-1}time^{-1}
d	0.01	IL-2 secretion rate by activated T cells	molecules cells^{-1}time^{-1}
e	0.01	IL-2 consumption rate by activated T cells and Tregs	cells^{-1}time^{-1}
f	1	IL-2 decay rate	time^{-1}
g	B	Natural death rate of naïve T cells and resting Tregs	time^{-1}
β	[0,∞)	Ag-stimulation of naïve T cells and resting Tregs	time^{-1}
γ	0.1	Treg-mediated suppression rate	cells^{-1}time^{-1}
ϵ	0.6	Proliferation rate ratio Treg/Tconv	–
N_0	4	Renewal rate of naïve T cells	cells time^{-1}
λ	0.006, 0.02	Relative renewal rate of resting Tregs and naïve T cells $\widehat{N}_0/N_0$	–
$\widehat{N}_0$	λ N_0	Renewal rate of resting Tregs	cells time^{-1}

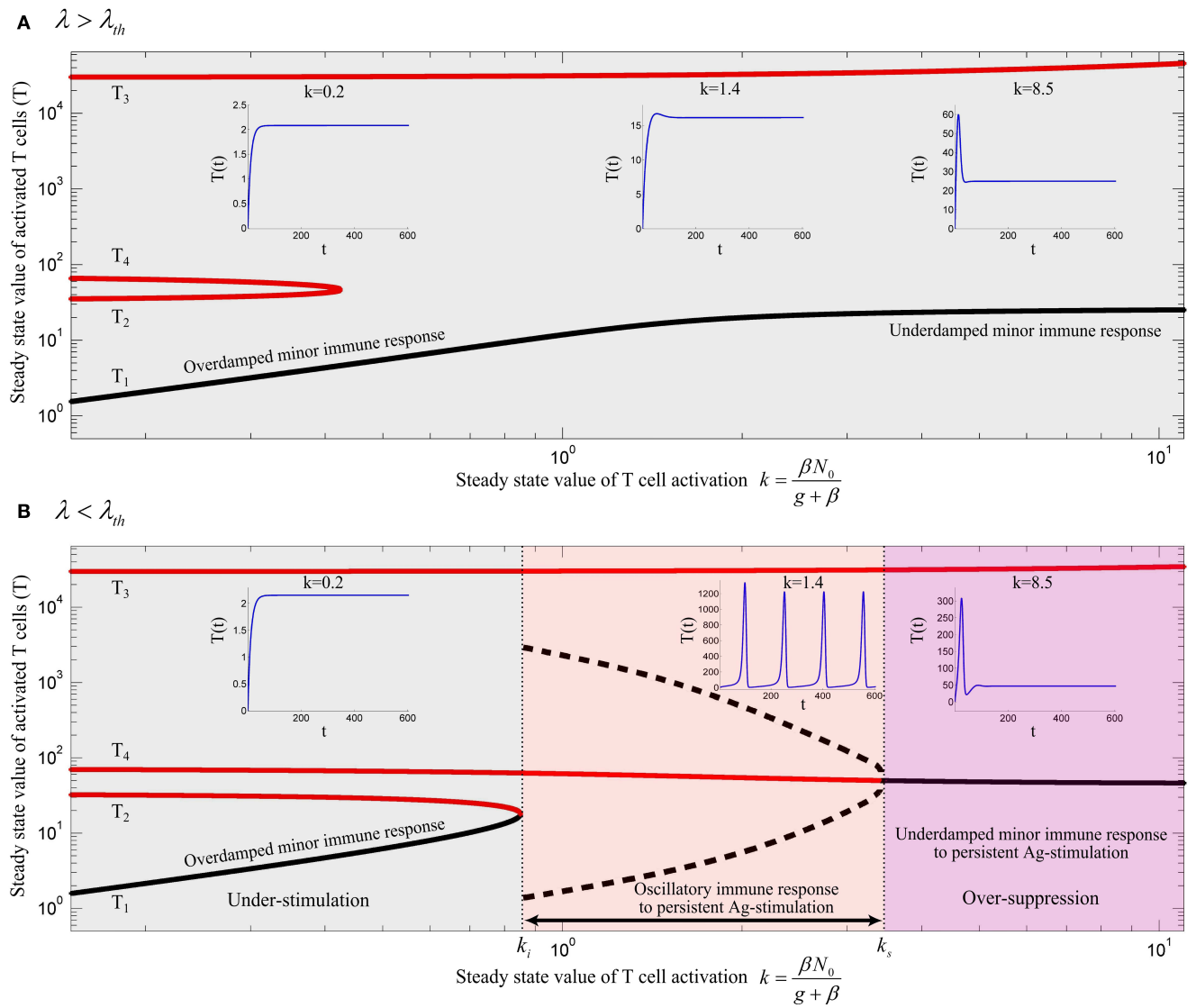


FIGURE 5 | Bifurcation diagram of model (13) using *k* as the bifurcation parameter with (A) $\lambda = 0.02$ and (B) $\lambda = 0.006$. Stable and unstable equilibrium points are drawn by black and red solid lines, respectively. Dashed black lines represent the stable limit cycles by showing the maximum and minimum populations of oscillating activated T cells for persistent *k*. Depending on the values of λ and *k*, an immune response is not initiated (gray), is initiated (red) or over-suppressed (magenta). With parameter values given in **Table 1**, the threshold becomes $\lambda_{th} = 0.01183$. The time-courses of the activated T cell population $T(t)$ were deduced from a numerical solution of model (13) with zero initial conditions and persistent β. The unstable negative equilibrium point (T_5) is not shown in the plots.

dependent on the availability of IL-2, sufficient activated T cells are required to secrete IL-2 and induce immune suppression. Therefore, activated T cells undergo the proliferation up to a level that sufficient IL-2 is available for Treg proliferation and subsequent immune suppression. In contrast, by facilitated Treg activation ($\widehat{k}$), less Treg proliferation is required for suppressing activated T cells which means that the dependency of immune suppression on proliferation of activated T cells decreases. Consequently, by increasing Ag-stimulation (β) in the range of $\beta_i < \beta < \beta_s$ (red region in **Figure 5B**), Treg activation ($\widehat{k}$) increases as well which results in a reduced maximum population of activated T cells (**Figure 5B**, dashed black line) and an increased frequency of oscillations. By further increasing Ag-stimulation to $\beta > \beta_s$ (magenta region in **Figure 5B**), Treg activation ($\widehat{k}$) completely prevent oscillating immune response.

In the same way, by increasing the relative homeostatic population of resting Tregs and naïve T cells ($\lambda > \lambda_{th}$), Treg activation increases up to a level that Treg suppression does not depend on the proliferative response of activated T cells. Thus, activated T cells are not able to enter the massive proliferation for any Ag-stimulation level, as shown in **Figure 5A**. Similar results were derived from a model that considers a nonlinear IL-2 dependent proliferation rate

of activated T cells and Tregs (see Nonlinear Proliferation Rate of Conventional and Regulatory T Cells in Appendix).

DISCUSSION

In this study, a model of the dynamic interplay between effector and regulatory immune responses was examined to investigate the requirements for the initiation of an immune response by Ag-stimulation. The model unifies several components developed in previous studies, such as IL-2 dependent proliferation of T cells (48), fratricide-induce programed cell death (37), IL-2 competition between activated T cell and activated Tregs (24), and Treg-mediated immune suppression (23, 24, 28). Homeostatic division of T cell compartments was not considered in the present study, such that the main renewal source of T cells in the absence of Ag-stimulation is the thymus. While the presented model is still simplifying the real situation in many aspects, the stability analysis revealed a number of reasonable results matching many experimental findings and allowing for an analysis of reasons for immune failure and autoimmunity.

The model predicts three qualitatively different immune responses depending on the level of antigenic stimulation. At first, a threshold stimulation β_i is required in order to get an immune response at all. Secondly, in a limited range of Ag-stimulation $\beta \in (\beta_i, \beta_s)$ an efficient immune response is induced. Tregs limit the duration of the immune response. If the antigen was cleared by the first immune response, further immune activity would be suppressed by Tregs. However, if the first peak of the immune response fails to clear the antigen, but keeps the antigen in the stimulation range $\beta_i < \beta < \beta_s$, the immune system attempts to clear the antigen with subsequent immune responses, which corresponds to the oscillatory solution depicted in **Figure 5B**. If the immune response failed to control the antigen spread, antigenic stimulation would be further increased to $\beta > \beta_s$, leading to the third class of immune responses. Tregs are over-stimulated and suppress immune activity. In this situation, a chronic persistence of the antigen would develop. Treg-mediated over-suppression of immune responses in chronic infections is well-established (reviewed in Ref. (49)). According to our model, depletion of resting Tregs restores the immune response by transiently decreasing λ and by this increasing β_s. This notion is consistent with the experimental model of chronic infections according to which depletion of Tregs results in the restoration of effector immune response and restriction of antigen spread (50, 51). A key feature of our model is that the immune response does not rely on a stable equilibrium point with a dominant population of activated T cells which is typically derived from existing bistable models. It rather relies on a transient response (or stable limit cycles in the case of persistent Ag-stimulation) which originates from T-cell-mediated suppression and IL-2 consumption by Tregs. Moreover, the role of Tregs in the chronic state of the immune response is not represented by available models.

According to our model, the qualitatively different immune responses and their requirements are dependent on the quality and quantity of Tconv and Treg clones responding to the Ag-stimulation. The proliferation rate of activated T cells, which depends on their avidity to the stimulating antigen determines the existence of an Ag-stimulation threshold (β_i) which is required for the initiation of an immune response. The absolute renewal rate of naïve T cells (N_0) adjusts the Ag-stimulation threshold β_i, which exists when the renewal rate of resting Tregs remains below a threshold value ($\lambda < \lambda_{th}$). Further Treg-associated parameters, namely the proliferation rate of Tregs (ϵ) and the Treg-mediated suppression rate (γ), also affect the existence and the level of the Ag-stimulation required for initiation (β_i) and over-suppression (β_s) of immune responses. By increasing the proliferative (ϵ) and suppressive (γ) activity of Tregs, β_i increases, whereas β_s decreases up to a level where the initiation of an immune response is completely impossible for any Ag-stimulation. Interestingly, when proliferation rate of activated Tregs exceeds the one for activated T cells ($\epsilon > 1$) a massive proliferation of activated T cells is still required for subsequent immune suppression by Tregs. Thus, IL-2 secretion by sufficiently large numbers of activated T cells is a strict requirement for immune suppression. Note also that without Tregs, a return to the homeostatic state is not possible, even when the antigen was cleared.

Considering all aforementioned parameters controlling the initiation of an immune response, is it beneficial for the immune system to completely avoid self reaction, or is there a benefit in allowing self reaction? Clearly, autoreactive T cells exist in the periphery of healthy individuals as a normal component of the T cell repertoire (12, 14, 52, 53). These cells respond to self-tissue destruction even in the presence of Tregs and without genetic predisposition to autoimmunity (15). Although the activation of autoreactive T cells has been shown to be involved in autoimmunity (12), several lines of evidences indicate that these cells are required for limiting self-destruction by supporting self-regenerative processes (54–56). In addition, the anti-tumor immune responses evoked by autoreactive T cells are beneficial (34, 57). Therefore, it seems unlikely that autoreactive T cells escaping from the thymus are simply a result of thymic selection error that can disturb self-tolerance under certain physiological conditions. Instead, these evidences imply that beneficial self reaction is allowed in the immune system. According to the mathematical model, immune reactions against self are only possible with a critical homeostatic population of autoreactive T cells (or sufficient renewal rate N_0) which is balanced by a proper number of Tregs ($\lambda < \lambda_{th}$) which corresponds to region (C) or (D) in **Figure 6**. Since the T cell repertoire is normally stimulated with an endogenous level of self-antigens in the periphery which does not evoke any self reaction, the Ag-stimulation threshold for initiating an immune response (β_i) should be sufficiently high in comparison to a typical nonself Ag-stimulation. According to our model, this is achieved by ensuring a low renewal rate (N_0) of low-avidity autoreactive T cells and a high, but balanced renewal rate of Tregs (high λ but lower than λ_{th}). In other words, according to **Figure 6**, by choosing N_0 close to k_i and higher value of k_i which is obtained by higher λ, a large Ag-stimulation threshold (β_i) for the initiation of immunity against self can be achieved.

Aging of the immune system, the so-called immunosenescence, is characterized by involution of thymus, decreased number of thymic output, contraction in T cell diversity, and disturbed T cell homeostasis which all result in attenuated immune function and susceptibility to infectious diseases and cancer in the elderly (58, 59). By decreasing thymic output, the homeostatic population of some T cell clones diminishes which leads to the creation of holes

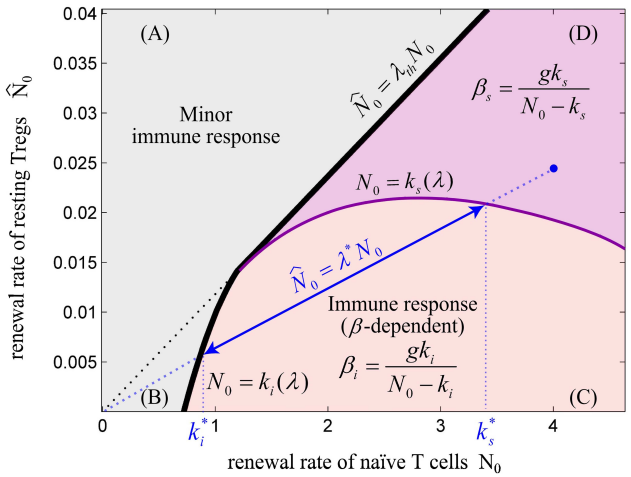


FIGURE 6 | The balance between renewal rate of naïve T cells and resting Tregs. The relative renewal rate of naïve T cells and resting Tregs ($\lambda = \widehat{N}_0/N_0$) determines the existence of an immune response. The initiation of an immune response requires a sufficient renewal rate of naïve T cells (N_0). **(A)** For $\lambda > \lambda_{th}$, the immune response does not exist for any value of N_0 and Ag-stimulation (β). **(B)** For $N_0 < k_i$, no immune response can be initiated for any value of β. **(C)** For $N_0 > k_i$, immune response can be initiated for $\beta > \beta_i$. In this regime, the Ag-stimulation that results in over-suppression of immune response (β_s) does not exist. **(D)** In this regime, immune response is initiated for $\beta > \beta_i$, and is over-suppressed for $\beta > \beta_s$. Note that k_i and k_s are dependent on the value of λ. For any point in the plane (e.g., blue point), the values of k_s^* and k_i^* are obtained by projecting the intersections of the line created by connecting the point to the origin (slope $= \lambda^*$) with the nonlinear curves $N_0 = k_i(\lambda)$ and $N_0 = k_s(\lambda)$ onto the N_0-axis. By decreasing λ^*, the effective range of T cell activation ($k_s^* - k_i^*$) or equivalently, the effective range of Ag-stimulation ($\beta_s^* - \beta_i^*$) which evoke immune response without over-suppression increases. We hypothesize that a healthy individual bears the potential to evoke self reaction and therefore its immune system is located in parametric regime **(C)** or **(D)**; however, higher self Ag-stimulation compared to nonself Ag-stimulation is required for initiating immune response due to low renewal rate of autoreactive T cells (N_0).

in the T cell repertoire (60). According to our model, a decreased renewal rate of a naïve T cell clone (N_0) *per se* could prevent an immune response or increase the Ag-stimulation level required for initiation of an immune response. In addition, as shown in many studies, the frequency of Tregs increases with age (61, 62) which results in a disturbed balance between the population of naïve T cells and resting Tregs (increased λ). In line with these results, in the mathematical model an increased λ prevents the initiation of an immune response corresponding to the age-related immune hyporesponsiveness in infection and cancer.

Based on the reasonable and physiologically realistic results that we could derive from the model, we dare to speculate about the self versus nonself concept emerging from the model. As mentioned before, the naïve T cells and resting Tregs are two major components of the immune reaction. The model does not distinguish self and nonself, but rather derives different responses to self and nonself from quantitative differences in the nature of Ag-stimulation. According to the model, by adjusting different parameters, different requirements in terms of Ag-stimulation level are found for the initiation of immune responses to self versus nonself. If the immune system responds according to a universal set of Ag-stimulation thresholds, regardless of whether the stimulus arises from self or nonself-antigens, a change of systemic parameters can lead to immune failure or autoimmunity. Self is no more considered as self if it exceeds an Ag-stimulation threshold determined by the stringency of central and peripheral tolerances. Similarly, nonself is considered as self if it does not properly stimulate the T cell repertoire. Autoimmunity might occur due to either a failure in tuning the Ag-stimulation threshold by the thymus that leads to unwanted self reaction in the periphery, or a chronic self Ag-stimulation in the periphery that leads to an oscillating self reaction and tissue destruction like in type 1 diabetes (63) and multiple sclerosis (64). Cancer or chronic infection would arise as the result of an imbalance in central and peripheral tolerances such as insufficient release of autoreactive T cells as well as high production or induction of Tregs that results in over-suppression of immune responses.

An early elegant mathematical modeling study analyzed a series of models to investigate self/nonself discrimination by T cells without explicitly considering suppressive Tregs (48). As a result of their critical assumption that memory cells accumulate in poor stimulatory conditions, the authors suggested that due to high stimulation by self antigens the lack of memory accumulation for T cell clones with high affinity to self accounts for self-tolerance. Also in our model, a high self Ag-stimulation ($\beta > \beta_s$ in **Figure 5B**) results in over-activation of Tregs and by this in over-suppression of self reaction. In both models an increased stimulation by self antigen would not lead to autoimmunity. The fact that autoreactive T cells do respond in the presence of Tregs when their stimulatory requirements are provided (15) makes it unlikely that this is the mechanism of self-tolerance induction. In the framework of our model, the view is supported that immune tolerance is induced by an increased stimulation threshold for self antigen and keeping self Ag-stimulation in a subcritical regime ($\beta < \beta_i$).

Undoubtedly, other mechanisms besides clonal deletion and Treg selection in the thymus also contribute to the fine tuning of the Ag-stimulation threshold required for initiation of immune reactions to self and nonself, such as anergy in the periphery (65) or activation threshold tuning in the thymus (66, 67). However, our simple model emphasizes the subtle balance between the generation of Tregs and autoreactive T cells which are both needed for beneficial autoimmunity. The model supports the view according to which self and nonself do not differ on a qualitative level. It is rather quantitative differences of the immune status and Ag-stimulation level that determine which molecule is treated as self or nonself.

ACKNOWLEDGMENTS

Michael Meyer-Hermann was supported by HFSP and BMBF within the GerontoSys initiative (projects GerontoMitoSys and GerontoShield). Sahamoddin Khailaie was supported by the Helmholtz International Graduate School for Infection Research.

REFERENCES

1. Babbitt BP, Allen PM, Matsueda G, Haber E, Unanue ER. Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature* (1985) **317**:359–61. doi:10.1038/317359a0

2. Guermonprez P, Valladeau J, Zitvogel L, Théry C, Amigorena S. Antigen presentation and T cell stimulation by dendritic cells. *Annu Rev Immunol* (2002) **20**:621–67. doi:10.1146/annurev.immunol.20.100301.064828
3. Yewdell JW, Haeryfar SM. Understanding presentation of viral antigens to CD8+ T cells in vivo: the key to rational vaccine design. *Annu Rev Immunol* (2005) **23**:651–82. doi:10.1146/annurev.immunol.23.021704.115702
4. Kanduc D, Stufano A, Lucchese G, Kusalik A. Massive peptide sharing between viral and human proteomes. *Peptides* (2008) **29**(10):1755–66. doi:10.1016/j.peptides.2008.05.022
5. Trost B, Kusalik A, Lucchese G, Kanduc D. Bacterial peptides are intensively present throughout the human proteome. *Self Nonself* (2010) **1**(1):71–4. doi:10.4161/self.1.1.9588
6. Wolfl M, Rutebemberwa A, Mosbruger T, Mao Q, Li H, Netski D, et al. Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire. *J Immunol* (2008) **181**(9):6435–46.
7. McCaughtry T, Hogquist K. Central tolerance: what have we learned from mice? *Semin Immunopathol* (2008) **30**:399–409. doi:10.1007/s00281-008-0137-0
8. Palmer E, Naeher D. Affinity threshold for thymic selection through a T-cell receptor-co-receptor zipper. *Nat Rev Immunol* (2009) **9**:207–13. doi:10.1038/nri2469
9. Rolland M, Nickle DC, Deng W, Frahm N, Brander C, Learn GH, et al. Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins. *PLoS One* (2007) **2**(9):e823. doi:10.1371/journal.pone.0000823
10. Frankild S, Boer RJD, Lund O, Nielsen M, Kesmir C. Amino acid similarity accounts for T cell cross-reactivity and for holes in the T cell repertoire. *PLoS One* (2008) **3**(3):e1831. doi:10.1371/journal.pone.0001831
11. Bouneaud C, Kourilsky P, Bousso P. Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion. *Immunity* (2000) **13**:829–40. doi:10.1016/S1074-7613(00)00080-7
12. Zehn D, Bevan M. T cells with low avidity for a tissue-restricted antigen routinely evade central and peripheral tolerance and cause autoimmunity. *Immunity* (2006) **25**:261–70. doi:10.1016/j.immuni.2006.06.009
13. Derbinski J, Kyewski B. How thymic antigen presenting cells sample the body's self-antigens. *Curr Opin Immunol* (2010) **22**(5):592–600. doi:10.1016/j.coi.2010.08.003
14. Bulek AM, Cole DK, Skowera A, Dolton G, Gras S, Madura F, et al. Structural basis for the killing of human beta cells by CD8(+) T cells in type 1 diabetes. *Nat Immunol* (2012) **13**(3):283–9. doi:10.1038/ni.2206
15. Enouz S, Carrié L, Merkler D, Bevan MJ, Zehn D. Autoreactive T cells bypass negative selection and respond to self-antigen stimulation during infection. *J Exp Med* (2012) **209**(10):1769–79. doi:10.1084/jem.20120905
16. Mueller DL. Mechanisms maintaining peripheral tolerance. *Nat Immunol* (2010) **11**(12):21–7. doi:10.1038/ni.1817
17. Hsieh C-S, Lee H-M, Lio C-WJ. Selection of regulatory T cells in the thymus. *Nat Rev Immunol* (2012) **12**(3):157–67. doi:10.1038/nri3155
18. Jordan MS, Boesteanu A, Reed AJ, Petrone AL, Holenbeck AE, Lerman MA, et al. Thymic selection of CD4+CD25+ regulatory T cells induced by an agonist self-peptide. *Nat Immunol* (2001) **2**:301–6. doi:10.1038/86302
19. Killebrew JR, Perdue N, Kwan A, Thornton AM, Shevach EM, Campbell DJ. A self-reactive TCR drives the development of Foxp3+ regulatory T cells that prevent autoimmune disease. *J Immunol* (2011) **187**(2):861–9. doi:10.4049/jimmunol.1004009
20. Kim JM, Rasmussen JP, Rudensky AY. Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice. *Nat Immunol* (2006) **8**:191–7. doi:10.1038/ni1428
21. Lahl K, Loddenkemper C, Drouin C, Freyer J, Arnason J, Eberl G, et al. Selective depletion of Foxp3+ regulatory T cells induces a scurfy-like disease. *J Exp Med* (2007) **204**(1):57–63. doi:10.1084/jem.20061852
22. León K, Peréz R, Lage A, Carneiro J. Modelling T-cell-mediated suppression dependent on interactions in multicellular conjugates. *J Theor Biol* (2000) **207**(2):231–54. doi:10.1006/jtbi.2000.2169
23. Carneiro J, Paixão T, Milutinovic D, Sousaa J, León K, Gardner R, et al. Immunological self-tolerance: lessons from mathematical modeling. *J Comput Appl Math* (2005) **184**(1):77–100. doi:10.1016/j.cam.2004.10.025
24. Burroughs NJ, Miguel Paz Mendes de Oliveira B, Adrego Pinto A. Regulatory T cell adjustment of quorum growth thresholds and the control of local immune responses. *J Theor Biol* (2006) **241**(1):134–41. doi:10.1016/j.jtbi.2005.11.010
25. Kim PS, Lee PP, Levy D. Modeling regulation mechanisms in the immune system. *J Theor Biol* (2007) **246**(1):33–69. doi:10.1016/j.jtbi.2006.12.012
26. Carneiro J, León K, Caramalho I, Van Den Dool C, Gardner R, Oliveira V, et al. When three is not a crowd: a cross regulation model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunol Rev* (2007) **216**(1):48–68. doi:10.1111/j.1600-065X.2007.00487.x
27. Burroughs NJ, Ferreira MF, Oliveira B, Pinto AA. Autoimmunity arising from bystander proliferation of T cells in an immune response model. *Math Comput Model* (2011) **53**(7):1389–93. doi:10.1016/j.mcm.2010.01.020
28. Almeida ARM, Amado IF, Reynolds J, Berges J, Lythe G, Molina-París C, et al. Quorum-sensing in CD4(+) T cell homeostasis: a hypothesis and a model. *Front Immunol* (2012) **3**:125. doi:10.3389/fimmu.2012.00125
29. León K, Peréz R, Lage A, Carneiro J. Three-cell interactions in T cell-mediated suppression? A mathematical analysis of its quantitative implications. *J Immunol* (2001) **166**(9):5356–65.
30. Thornton AM, Shevach EM. Suppressor effector function of CD4+CD25+ immunoregulatory T cells is antigen nonspecific. *J Immunol* (2000) **164**(1):183–90.
31. Pacholczyk R, Kern J, Singh N, Iwashima M, Kraj P, Ignatowicz L. Nonself-antigens are the cognate specificities of Foxp3+ regulatory T cells. *Immunity* (2007) **27**(3):493–504. doi:10.1016/j.immuni.2007.07.019
32. Hofmann M, Radsak M, Rechtsteiner G, Wiemann K, Günder M, Bien-Gräter U, et al. T cell avidity determines the level of CTL activation. *Eur J Immunol* (2004) **34**(7):1798–806. doi:10.1002/eji.200425088
33. Zehn D, Lee SY, Bevan MJ. Complete but curtailed T-cell response to very low-affinity antigen. *Nature* (2009) **458**(7235):211–4. doi:10.1038/nature07657
34. Hebeisen M, Rufer N, Oberle SG, Speiser DE, Zehn D. Signaling mechanisms that balance anti-viral, auto-reactive, and anti-tumor potential of low affinity T cells. *J Clin Cell Immunol* (2012) **S12**(3). doi:10.4172/2155-9899.S12-003
35. Bourgeois C, Hao Z, Rajewsky K, Potocnik AJ, Stockinger B. Ablation of thymic export causes accelerated decay of naïve CD4 T cells in the periphery because of activation by environmental antigen. *Proc Natl Acad Sci U S A* (2008) **105**(25):8691–6. doi:10.1073/pnas.0803732105
36. Takada K, Jameson SC. Naive T cell homeostasis: from awareness of space to a sense of place. *Nat Rev Immunol* (2009) **9**(12):823–32. doi:10.1038/nri2657
37. Callard RE, Stark J, Yates AJ. Fratricide: a mechanism for T memory-cell homeostasis. *Trends Immunol* (2003) **24**(7):370–5. doi:10.1016/S1471-4906(03)00164-9
38. Belkaid Y. Regulatory T cells and infection: a dangerous necessity. *Nat Rev Immunol* (2007) **7**:875–88. doi:10.1038/nri2189
39. Zhou Y. Regulatory T cells and viral infections. *Front Biosci* (2008) **13**:1152–70. doi:10.2741/2752
40. Kretschmer K, Apostolou I, Jaeckel E, Khazaie K, von Boehmer H. Making regulatory T cells with defined antigen specificity: role in autoimmunity and cancer. *Immunol. Rev* (2006) **212**:163–9. doi:10.1111/j.0105-2896.2006.00411.x
41. Klages K, Mayer CT, Lahl K, Loddenkemper C, Teng MW, Ngiow SF, et al. Selective depletion of Foxp3+ regulatory T cells improves effective therapeutic vaccination against established melanoma. *Cancer Res* (2010) **70**(20):7788–99. doi:10.1158/0008-5472.CAN-10-1736
42. Vignali DA, Collison LW, Workman CJ. How regulatory T cells work. *Nat Rev Immunol* (2008) **8**(7):523–32. doi:10.1038/nri2343
43. Thornton AM, Korty PE, Tran D, Wohlfert EA, Murray PE, Belkaid Y, et al. Expression of Helios, an Ikaros transcription factor family member, differentiates thymic-derived from peripherally induced Foxp3+ T regulatory cells. *J Immunol* (2010) **184**:3433–41. doi:10.4049/jimmunol.0904028
44. Fontenot JD, Rasmussen JP, Gavin MA, Rudensky AY. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nat Immunol* (2005) **6**:1142–51. doi:10.1038/ni1263
45. Zou T, Caton AJ, Koretzky GA, Kambayashi T. Dendritic cells induce regulatory T cell proliferation through antigen-dependent and -independent interactions. *J Immunol* (2010) **185**(5):2790–9. doi:10.4049/jimmunol.0903740
46. Busse D, de la Rosa M, Hobiger K, Thurley K, Flossdorf M, Scheffold A, et al. Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments. *Proc Natl Acad Sci U S A* (2010) **107**(7):3058–63. doi:10.1073/pnas.0812851107

47. Sakaguchi S. Naturally arising Foxp3-expressing CD25+CD4+ regulatory T cells in immunological tolerance to self and non-self. *Nat Immunol* (2005) **6**:345–52. doi:10.1038/ni1178
48. de Boer RJ, Hogeweg P. Self-nonself discrimination due to immunological non-linearities: the analysis of a series of models by numerical methods. *Math Med Biol* (1987) **4**(1):1–32. doi:10.1093/imammb/4.1.1
49. Li S, Gowans EJ, Chougnet C, Plebanski M, Dittmer U. Natural regulatory T cells and persistent viral infection. *J Virol* (2008) **82**(1):21–30. doi:10.1128/JVI.01768-07
50. Dietze KK, Zelinskyy G, Gibbert K, Schimmer S, Francois S, Myers L, et al. Transient depletion of regulatory T cells in transgenic mice reactivates virus-specific CD8 + T cells and reduces chronic retroviral set points. *Proc Natl Acad Sci U S A* (2011) **108**(6):2420–5. doi:10.1073/pnas.1015148108
51. Keynan Y, Card CM, McLaren PJ, Dawood MR, Kasper K, Fowke KR. The role of regulatory T cells in chronic and acute viral infections. *Clin Infect Dis* (2008) **46**(7):1046–52. doi:10.1086/529379
52. von Herrath MG, Dockter J, Oldstone MB. How virus induces a rapid or slow onset insulin-dependent diabetes mellitus in a transgenic model. *Immunity* (1994) **1**:231–42. doi:10.1016/1074-7613(94)90101-5
53. Turner MJ, Jellison ER, Lingenheld EG, Puddington L, Lefrançois L. Avidity maturation of memory CD8 T cells is limited by self-antigen expression. *J Exp Med* (2008) **205**(8):1859–68. doi:10.1084/jem.20072390
54. Barouch R, Schwartz M. Autoreactive T cells induce neurotrophin production by immune and neural cells in injured rat optic nerve: implications for protective autoimmunity. *FASEB J* (2002) **16**(10):1304–6. doi:10.1096/fj.01-0467fje
55. Hofstetter HH, Sewell DL, Liu F, Sandor M, Forsthuber T, Lehmann PV, et al. Autoreactive T cells promote post-traumatic healing in the central nervous system. *J Neuroimmunol* (2003) **134**(1):25–34. doi:10.1016/S0165-5728(02)00358-2
56. Wekerle H, Hohlfeld R. Beneficial brain autoimmunity? *Brain* (2010) **133**(8):2182–4. doi:10.1093/brain/awq206
57. Baitsch L, Fuertes-Marraco SA, Legat A, Meyer C, Speiser DE. The three main stumbling blocks for anticancer T cells. *Trends Immunol* (2012) **33**(7):364–72. doi:10.1016/j.it.2012.02.006
58. Goronzy JJ, Weyand CM. Understanding immunosenescence to improve responses to vaccines. *Nat Immunol* (2013) **14**:428–36. doi:10.1038/ni.2588
59. Vadasz Z, Haj T, Kessel A, Toubi E. Age-related autoimmunity. *BMC Med* (2013) **11**:94. doi:10.1186/1741-7015-11-94
60. Yager EJ, Ahmed M, Lanzer K, Randall TD, Woodland DL, Blackman MA. Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J Exp Med* (2008) **205**(3):711–23. doi:10.1084/jem.20071140
61. Sakaguchi S, Miyara M, Costantino CM, Hafler DA. FOXP3+ regulatory T cells in the human immune system. *Nat Rev Immunol* (2010) **10**:490–500. doi:10.1038/nri2785
62. Raynor J, Lages CS, Shehata H, Hildeman DA, Chougnet CA. Homeostasis and function of regulatory T cells in aging. *Curr Opin Immunol* (2012) **24**(4):482–7. doi:10.1016/j.coi.2012.04.005
63. von Herrath M, Sanda S, Herold K. Type 1 diabetes as a relapsing-remitting disease? *Nat Rev Immunol* (2007) **7**(12):988–94. doi:10.1038/nri2192
64. Nylander A, Hafler DA. Multiple sclerosis. *J Clin Invest* (2012) **122**(4):1180–8. doi:10.1172/JCI58649
65. Lechler R, Chai JG, Marelli-Berg F, Lombardi G. The contributions of T-cell anergy to peripheral T-cell tolerance. *Immunology* (2001) **103**(3):262–9. doi:10.1046/j.1365-2567.2001.01250.x
66. Grossman Z, Singer A. Tuning of activation thresholds explains flexibility in the selection and development of T cells in the thymus. *Proc Natl Acad Sci U S A* (1996) **93**(25):14747–52. doi:10.1073/pnas.93.25.14747
67. Scherer A, Noest A, de Boer RJ. Activation-threshold tuning in an affinity model for the T-cell repertoire. *Proc Biol Sci* (2004) **271**:609–16. doi:10.1098/rspb.2003.2653
68. Banz A, Pontoux C, Papiernik M. Modulation of Fas-dependent apoptosis: a dynamic process controlling both the persistence and death of CD4 regulatory T cells and effector T cells. *J Immunol* (2002) **169**(2):750–7.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 August 2013; accepted: 06 December 2013; published online: 26 December 2013.

*Citation: Khailaie S, Bahrami F, Janahmadi M, Milanez-Almeida P, Huehn J and Meyer-Hermann M (2013) A mathematical model of immune activation with a unified self-nonself concept. Front. Immunol. **4**:474. doi: 10.3389/fimmu.2013.00474*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Khailaie, Bahrami, Janahmadi, Milanez-Almeida, Huehn and Meyer-Hermann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

A.1. STEADY STATE ANALYSIS OF MODEL (1)

The equilibrium points of model (1) can be obtained from the following:

$$\text{Equilibrium points} = \begin{cases} (T_1, I_1) = (0,0) \\ (T_2, I_2) = \left(\dfrac{bf}{ad-be}, \dfrac{b}{a}\right) \end{cases} \tag{A1}$$

For a positive nontrivial equilibrium point (T_2, I_2), we have to assume that:

$$ad - be > 0 \tag{A2}$$

The stability of equilibrium points can be determined from the sign of real part of eigenvalues of Jacobian matrix (J). An equilibrium point is stable if all the eigenvalues of J evaluated at the equilibrium point have negative real parts, and it is unstable if at least one of the eigenvalues has a positive real part.

$$\text{Jacobian Matrix J} = \begin{bmatrix} aI - b & aT \\ d - eI & -eT - f \end{bmatrix} \tag{A3}$$

$$\begin{aligned} \text{Characteristic Equation } Q(\lambda) = det\left\{\begin{bmatrix} \lambda - aI + b & -aT \\ -d + eI & \lambda + eT + f \end{bmatrix}\right\} \\ = \lambda^2 + \left[eT + f + b - aI\right]\lambda + \left[-afI + beT + bf - adT\right] = 0 \end{aligned} \tag{A4}$$

The eigenvalues ($\lambda_{1,2}$) of J for trivial equilibrium point (T_1, I_1) are obtained by solving the characteristic equation (A4):

$$\lambda_{1,2}\big|_{(T_1,I_1)} : Q(\lambda)\big|_{(T_1,I_1)} = \lambda^2 + (f+b)\lambda + bf = 0 \tag{A5}$$

By checking Routh–Hurwitz stability Criterion (RHC) it can be easily confirmed that the eigenvalues have negative real parts since all the coefficients of polynomial $Q(\lambda)$ are positive, and hence, the trivial equilibrium point (T_1, I_1) is locally stable. For stability of nontrivial equilibrium point, the characteristic equation (A4) is evaluated and solved in (T_2, I_2):

$$\lambda_{1,2}\big|_{(T_2,I_2)} : Q(\lambda)\big|_{(T_2,I_2)} = \lambda^2 + \left(\frac{afd}{ad-be}\right)\lambda - bf = 0 \tag{A6}$$

With the assumption (A2), the coefficient of λ is positive. The sign of $Q(\lambda)$ coefficients change only once and hence, there exists one positive eigenvalue. Therefore, the nontrivial equilibrium point (T_2, I_2) is a saddle point and unstable.

A.2. STEADY STATE ANALYSIS OF MODEL (4)

The equilibrium points of model (4) with definition of T cell activation ($k(t)$) given in equation (5) can be obtained from the following

$$N = \frac{N_0}{g+\beta} = \frac{k}{\beta} \tag{A7}$$

$$I = \frac{dT}{eT+f} = \frac{bT-k}{aT} \tag{A8}$$

$$(ad-be)\,T^2 + (ek - bf)\,T + fk = 0 \tag{A9}$$

By keeping the assumption (A2), if the coefficient of T in (A9) is negative, the equilibrium points (T), if exist, will be positive:

$$ek - bf < 0 \rightarrow k < \frac{bf}{e} \tag{A10}$$

Otherwise, the equilibrium points will be negative. According to equation (A8), for $T \geq 0, I \geq 0$. Now, let's check the condition for existence of equilibrium points:

$$\Delta = 0 \rightarrow \left(ek + bf\right)^2 - 4afdk = \underbrace{\left(ek + bf + 2\sqrt{afdk}\right)}_{>0}\underbrace{\left(ek + bf - 2\sqrt{afdk}\right)}_{=0} = 0 \tag{A11}$$

$$\left(ek + bf - 2\sqrt{afdk}\right) = 0 \rightarrow k_+ = \frac{adf}{e^2}\left(1 + \sqrt{1 - \frac{be}{ad}}\right)^2, k_- = \frac{adf}{e^2}\left(1 - \sqrt{1 - \frac{be}{ad}}\right)^2 \tag{A12}$$

The model does not have any equilibrium points for $k_- < k < k_+$, and two equilibrium points otherwise. It can be verified that by keeping assumption (A2), we always have:

$$0 < k_- < \frac{bf}{e} < k_+ \tag{A13}$$

Therefore, for $0 < k < k_-$, condition (A10) is satisfied and the model has two positive equilibrium points and for $k > k_+$, condition (A10) is not satisfied and model has two negative equilibrium points. Let's assume that the model has two positive equilibrium points ($\Delta > 0$ and $0 < k < k_-$). In the following, the linear stability of equilibrium points is analyzed:

$$\text{Jacobian Matrix J} = \begin{bmatrix} -g - \beta & 0 & 0 \\ \beta & aI - b & aT \\ 0 & d - eI & -eT - f \end{bmatrix} \tag{A14}$$

$$\begin{aligned} \text{Characteristic Equation } Q(\lambda) &= det\left\{\begin{bmatrix} \lambda + g + \beta & 0 & 0 \\ -\beta & \lambda - aI + b & -aT \\ 0 & -d + eI & \lambda + eT + f \end{bmatrix}\right\} \\ &= \left[\lambda + \left(g + \beta\right)\right]\underbrace{\left[\left(\lambda - aI + b\right)\left(\lambda + eT + f\right) + aT\left(eI - d\right)\right]}_{Q^*} = 0 \end{aligned} \tag{A15}$$

The model has one negative eigenvalue $\lambda = -(g + \beta)$ for all equilibrium points. For the other two remaining eigenvalues, polynomial Q^* has to be checked for existence of positive eigenvalue.

$$Q^* = \lambda^2 + \underbrace{\left[eT + f + b - aI\right]}_{U}\lambda + \underbrace{\left[-afI + beT + bf - adT\right]}_{V} = 0 \tag{A16}$$

From equation (A8) it can be easily verified that $b - aI > 0$ and hence, coefficient U is positive. Therefore, the stability depends on the sign of coefficient V.

$$V = -afI + beT + bf - adT \xrightarrow{I = \frac{dT}{eT + f}} V = -\frac{afd}{eT + f}T - \left(ad - be\right)T + bf \tag{A17}$$

$$V_1 = TV = -\frac{afd}{eT + f}T^2 - \underbrace{\left[\left(ad - be\right)T^2 + \left(ek - bf\right)T + fk\right]}_{\text{According to (A9)} \rightarrow = 0} + k\left(eT + f\right) \tag{A18}$$

$$V_2 = \frac{1}{k\left(eT + f\right)}V_1 = -\frac{af}{kd}\left(\frac{dT}{eT + f}\right)^2 + 1 \xrightarrow{I = \frac{dT}{eT + f}} V_2 = -\frac{af}{kd}I^2 + 1 \tag{A19}$$

According to equations (A8) and (A9), the equilibrium values of I are:

$$I_2 = \frac{1}{2}\left(\frac{\left(ek + bf\right) + \sqrt{\Delta}}{af}\right), T_2 = \frac{k}{b - aI_2}, N_2 = \frac{N_0}{g + \beta} \tag{A20}$$

$$I_1 = \frac{1}{2}\left(\frac{\left(ek + bf\right) - \sqrt{\Delta}}{af}\right), T_1 = \frac{k}{b - aI_1}, N_1 = \frac{N_0}{g + \beta} \tag{A21}$$

where $I_2 > I_1$ and:

$$\Delta = \left(ek + bf\right)^2 - 4fkad = \left(ek - bf\right)^2 - 4fk\left(ad - be\right) \tag{A22}$$

Next, the sign of V_2, which is similar to V_1 and V, has to be checked in the equilibrium points. For the larger equilibrium point (T_2, I_2):

$$\begin{aligned} V_2\left|_{I_2}\right. &= -\frac{1}{2fkad}\left[e^2k^2 + 2ekbf + ek\sqrt{\Delta} + b^2f^2 + bf\sqrt{\Delta} - 4fkad\right] \\ &= -\frac{1}{2fkad}\underbrace{\left[\Delta + \left(ek + bf\right)\sqrt{\Delta}\right]}_{+} < 0 \end{aligned} \tag{A23}$$

V_2, V_1, and V are negative for (T_2, I_2). Therefore, the model in this equilibrium point has a positive eigenvalue and it is locally unstable. For (T_1, I_1),

$$V_2\left|_{I_1}\right. = -\frac{\sqrt{\Delta}}{2fkad}\left[\sqrt{\Delta} - \left(ek + bf\right)\right] \overset{\text{According to (A22)}}{=\!=} -\frac{\sqrt{\Delta}}{2fkad}\underbrace{\left[\sqrt{\left(ek + bf\right)^2 - 4fkad} - \left(ek + bf\right)\right]}_{-} > 0 \tag{A24}$$

V_2, V_1, and V are positive for (T_1, I_1). Therefore, all the eigenvalues of the model in this equilibrium point are negative and it is locally stable.

Next, let's assume that $k > k_+$ which means that the equilibrium points exist and the steady state values of T are negative, whereas the equilibrium values of I is positive. According to equation (A8), coefficient U in equation (A16) is negative since:

$$T = \frac{k}{b - aI} < 0 \rightarrow b - aI < 0, I = \frac{dT}{eT + f} > 0 \xrightarrow{dT<0} eT + f < 0 \tag{A25}$$

Therefore, the model at least has one positive eigenvalue in the equilibrium points. Let's check the sign of coefficient V in the equilibrium points:

$$V = -afI + beT + bf - adT \overset{T=\frac{k}{b-aI}}{=\!=} -f\left(aI - b\right) + \left(ad - be\right)\frac{k}{aI - b} \tag{A26}$$

$$V_2 = \left(aI - b\right)V \overset{aI-b>0}{=\!=} -f(aI - b)^2 + \left(ad - be\right)k \overset{aI-b=-\frac{k}{T}}{=\!=} -f\frac{k^2}{T^2} + \left(ad - be\right)k$$

$$V_3 = \frac{T^2}{fk^2}V_2 = -1 + \left(\frac{ad - be}{fk}\right)T^2 \tag{A27}$$

The sign of V_3 in equation (A27) which is the same as the sign of V in equation (A16) can be determined by substituting T with its equilibrium values from equation (A9):

$$V_3\left|_{T_2}\right. = \frac{1}{4\left(ad - be\right)fk}\underbrace{\left[2\Delta + 2\sqrt{\Delta}\left(bf - ek\right)\right]}_{\text{sign?}} \tag{A28}$$

According to equation (A13) and reminding that $k > k_+$, it can be verified that V_3 or equivalently V is positive. Therefore, the equilibrium point (T_2, I_2) has one positive eigenvalue and is unstable. For (T_1, I_1),

$$V_3\left|_{T_1}\right. = \frac{1}{4\left(ad - be\right)fk}\underbrace{\left[2\Delta + 2\sqrt{\Delta}\underbrace{\left(ek - bf\right)}_{+}\right]}_{+} \tag{A29}$$

V_3 and equivalently, V are positive and therefore, the equilibrium point (T_1, I_1) has one positive eigenvalue and it is unstable.

A.3. STEADY STATE ANALYSIS OF MODEL (10)

The equilibrium points of model (10) for $\beta = 0$ can be obtained from the following

$$(T_1,\ I_1,\ N_1) = \left(0,\ 0,\ \frac{N_0}{g}\right) \tag{A30}$$

$$N_{2,3} : N_{2,3} = \frac{N_0}{g} \tag{A31}$$

$$T_{2,3} : ceT^2 + (cf + be - ad)T + bf = 0, \begin{cases} T_3 = -\frac{1}{2ce}\left[-\left(cf + be - ad\right) + \sqrt{\Delta}\right] \\ T_2 = -\frac{1}{2ce}\left[-\left(cf + be - ad\right) - \sqrt{\Delta}\right] \\ \Delta = (cf + be - ad)^2 - 4cfbe \end{cases} \tag{A32}$$

$$I_{2,3} : I_{2,3} = \frac{cT_{2,3} + b}{a} \tag{A33}$$

According to equation (A32), the nontrivial equilibrium points, if available, are positive only if:

$$cf + be - ad < 0 \tag{A34}$$

The condition of the existence of positive equilibrium points can be obtained from equation (A32):

$$\Delta = \left(cf + be - ad\right)^2 - 4cebf = f^2c^2 - 2f\,(ad + be)\,c + (ad - be)^2 \geq 0 \tag{A35}$$

According to equation (A35), the nontrivial equilibrium points exist only if:

$$\{c < c_-\} \cup \{c > c_+\} \tag{A36}$$

where

$$\left[c_- = \frac{\left(\sqrt{ad} - \sqrt{be}\right)^2}{f}\right] < \left[\frac{ad - be}{f} = \frac{\left(\sqrt{ad} + \sqrt{be}\right)\left(\sqrt{ad} - \sqrt{be}\right)}{f}\right] < \left[c_+ = \frac{\left(\sqrt{ad} + \sqrt{be}\right)^2}{f}\right] \tag{A37}$$

Therefore, according to conditions (A34) and (A36) and inequality (A37), the two positive equilibrium points exist only if:

$$0 < c < c_- \tag{A38}$$

Next, we assume that the condition (A38) is satisfied, and we analyze the stability of the equilibrium points:

$$\text{Jacobian Matrix J} = \begin{bmatrix} -g - \beta & 0 & 0 \\ \beta & aI - b - 2cT & aT \\ 0 & d - eI & -eT - f \end{bmatrix} \tag{A39}$$

Characteristic Equation:

$$\begin{aligned} Q(\lambda) &= det\left\{\begin{bmatrix} \lambda + g + \beta & 0 & 0 \\ -\beta & \lambda - aI + b + 2cT & -aT \\ 0 & -d + eI & \lambda + eT + f \end{bmatrix}\right\} \\ &= \left[\lambda + g + \beta\right]\left[\lambda^2 + \underbrace{\left[eT + f + b - aI + 2cT\right]}_{U}\lambda + \underbrace{\left[-afI + 2ceT^2 + beT + bf + 2cfT - adT\right]}_{V}\right] \end{aligned} \tag{A40}$$

All the equilibrium points have a negative eigenvalue $\lambda = -g - \beta$. For the sign of other eigenvalues, the sign of coefficients of the characteristic equation has to be checked in equilibrium points. These coefficients are positive for the trivial equilibrium point and

therefore, all eigenvalues have negative real value. Hence, the trivial equilibrium point is stable. For stability analysis of nontrivial equilibrium points, the sign of U and V in equation (A40) has to be analyzed:

$$U = eT + f + 2cT + (b - aI) \overset{(A33):\ b-aI=-cT}{=\!=} eT + f + cT > 0 \tag{A41}$$

Coefficient U is positive for the positive equilibrium points. Therefore, their stability depends on the sign of V in equation (A40):

$$V = f(b - aI) + 2ceT^2 + beT + 2cfT - adT \overset{(A33):\ b-aI=-cT}{=\!=} 2ceT^2 + (be - ad)T + cfT \tag{A42}$$

$$= ceT^2 + \underbrace{\left[ceT^2 + \left(cf + be - ad\right)T + bf\right]}_{\text{According to } (A32):\ =0} - bf = ceT^2 - bf \tag{A43}$$

$$V\big|_{T_3} = \frac{1}{2ce}\left[\Delta - \underbrace{\left(cf + be - ad\right)}_{\text{According to } (A34):\ <0}\sqrt{\Delta}\right] > 0 \tag{A44}$$

The sign of V for the larger equilibrium point is positive and hence, this equilibrium point is stable.

$$V|_{T_2} = \underbrace{\frac{\sqrt{\Delta}}{2ce}\left[\sqrt{\Delta} + \underbrace{\left(cf + be - ad\right)}_{\text{According to } (A34):\ <0}\right]}_{\text{According to } (A32):<0} < 0 \tag{A45}$$

The sign of V for the smaller equilibrium point is negative and therefore, this equilibrium point is unstable. The equilibrium points of the model with $\beta > 0$ and by considering $\beta N(t) = k(t)$ are obtained by solving:

$$-ceT^3 - \left(cf + be - ad\right)T^2 + \left(ek - bf\right)T + kf = 0 \tag{A46a}$$

$$I = \frac{dT}{eT + f} \tag{A46b}$$

By keeping the assumptions (A2) and (A38), the equation (A46a) has either one positive real equilibrium point or three positive real equilibrium points depending on the value of k. The stability of equilibrium points that are obtained from (A46a) and (A46b) by varying the value of k is hard to be checked analytically; instead, it is analyzed numerically by solving equation (A40).

A.4. STEADY STATE ANALYSIS OF MODEL (13)

The equilibrium points of model (13) for $\beta N(t) = k(t) = 0$ can be obtained from the following

$$(T_1,\ I_1,\ R_1,\ N_1) = \left(0,\ 0,\ 0,\ \frac{N_0}{g}\right) \tag{A47}$$

$$T_{2,3} : ceT_{2,3}^2 + (cf - ad + be)T_{2,3} + bf = 0, I_{2,3} = \frac{dT_{2,3}}{eT_{2,3} + f}, R_{2,3} = 0, N_{2,3} = \frac{N_0}{g} \tag{A48}$$

$$T_4 = \frac{b(eR_5 + f)}{\epsilon ad - be}, I_4 = \frac{b}{\epsilon a}, R_4 = \left(\frac{b}{\epsilon}\right)\left(\frac{-\epsilon(cf + be - ad) + be(\epsilon - 1) - \epsilon(\epsilon ad - be)}{cbe + \gamma(\epsilon ad - be)}\right), N_4 = \frac{N_0}{g} \tag{A49}$$

$$T_5 = 0,\ I_5 = \frac{b}{\epsilon a}, R_5 = \frac{-f}{e}, N_5 = \frac{N_0}{g} \tag{A50}$$

The sign of the equilibrium point T_4 changes by changing the Treg-associated parameters (ϵ, γ). T_4 is positive if

$$\epsilon ad - be > 0, R > -\frac{f}{e} \tag{A51}$$

or

$$\epsilon ad - be < 0, R < -\frac{f}{e} \tag{A52}$$

The physiologically relevant regime of the model occurs by satisfying the condition (A51) which means both T_4 and R_4 are positive. The equilibrium points of model (13) for $\beta\ N(t) = k(t) \neq 0$ can be obtained from the following

$$N = \frac{\beta N_0}{g + \beta} \tag{A53}$$

$$R = \frac{-\lambda k}{\epsilon aI - b} \tag{A54}$$

$$T = \frac{-I\left(-e\lambda k + f\epsilon aI - bf\right)}{(\epsilon aI - b)\,(-d + eI)} \tag{A55}$$

$$P_5 I^5 + P_4 I^4 + P_3 I^3 + P_2 I^2 + P_1 I + P_0 = 0 \tag{A56}$$

where

$$P_5 = a^3 f\epsilon^2 e \tag{A57}$$

$$P_4 = -bf\epsilon^2 a^2 e + cf^2\epsilon^2 a^2 - k\epsilon^2 a^2 e^2 - a^3 f\epsilon^2 d - 2\,a^2 f\epsilon\ be - a^2 e^2\lambda\ k\epsilon \tag{A58}$$

$$\begin{aligned} P_3 = {} & be^2\lambda\ k\epsilon\ a - 2\,cf^2\epsilon\ ab + ae^2\lambda\ kb + 2\,a^2 f\epsilon\ bd + \gamma\ \lambda\ kf\epsilon\ ae + ab^2 fe + bf\epsilon^2 a^2 d \\ & + 2\,k\epsilon^2 a^2 de - 2\,ce\lambda\ kf\epsilon\ a + 2\,b^2 f\epsilon\ ae + a^2 e\lambda\ k\epsilon\ d + 2\,k\epsilon\ abe^2 \end{aligned} \tag{A59}$$

$$\begin{aligned} P_2 = {} & 2\,ce\lambda\ kbf - 4\,k\epsilon\ abde - \gamma\ \lambda\ kbfe - b^2 e^2\lambda\ k - kb^2 e^2 - be\lambda\ k\epsilon\ ad - b^3 fe - ae\lambda\ kbd + cb^2 f^2 \\ & + ce^2\lambda^2 k^2 - ab^2 fd - k\epsilon^2 a^2 d^2 - \gamma\ \lambda^2 k^2 e^2 - 2\,b^2 f\epsilon\ ad - \gamma\ \lambda\ kf\epsilon\ ad \end{aligned} \tag{A60}$$

$$P_1 = \gamma\ \lambda^2 k^2 ed + b^2 e\lambda\ kd + 2\,kb^2 de + 2\,k\epsilon\ abd^2 + b^3 fd + \gamma\ \lambda\ kbfd \tag{A61}$$

$$P_0 = -kb^2 d^2 \tag{A62}$$

A.5. NONLINEAR PROLIFERATION RATE OF CONVENTIONAL AND REGULATORY T CELLS

In the models (1), (4), (10) and (13) it is assumed that proliferation rate of Tconvs and Tregs is a linear function of IL-2. This simplifying assumption is made in order to allow parametric stability analysis of the model in a closed form and to find explicitly the dependency between parametric variations and topological changes of the model. Here, we show that the simplifying assumption does not affect the three regimes of qualitative immune responses that could be derived from the model. The linear IL-2-dependent proliferation rate is replaced with a nonlinear function of IL-2, named $\Phi\ (I)$ in models (1), (4), and (13):

$$\begin{cases} \dfrac{\mathrm{d}T}{\mathrm{d}t} = \Phi(I)\,T - bT + k \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = dT - eIT - fI \end{cases} \tag{A63}$$

where $\Phi\ (I)$ is considered as a Hill-function of IL-2

$$\Phi(I) = a\frac{I^n}{h^n + I^n} \tag{A64}$$

The models (1) and (A63) are compared by steady state analysis. The equilibrium points of the modified model (A63) (with $k = \frac{\beta N_0}{g+\beta} = 0$) are

$$(T_1,\ I_1) = (0,\ 0) \tag{A65}$$

$$(T_2, I_2) = \left(\frac{fI_2}{d - eI_2}, h\left(\frac{b}{a - b}\right)^{\frac{1}{n}}\right) \tag{A66}$$

The nontrivial equilibrium point (T_2, I_2) is positive and biologically meaningful only if

$$(a - b)d^n - be^n h^n > 0 \tag{A67}$$

The local stability of the equilibrium points can be determined by obtaining the eigenvalues from the characteristic equation:

$$\text{Characteristic Equation } Q(\lambda) = det\left\{\begin{bmatrix} \lambda - a\frac{I^n}{h^n+I^n} + b & -aT\frac{nh^nI^{n-1}}{(h^n+I^n)^2} \\ -d+eI & \lambda + eT + f \end{bmatrix}\right\} \tag{A68}$$

$$= \lambda^2 + \left[eT + f - a\frac{I^n}{h^n+I^n} + b\right]\lambda + \left[\left(-a\frac{I^n}{h^n+I^n} + b\right)(eT+f) + aT\frac{nh^nI^{n-1}}{(h^n+I^n)^2}(-d+eI)\right] = 0 \tag{A69}$$

By checking Routh–Hurwitz stability Criterion (RHC) it can be easily confirmed that the eigenvalues have negative real parts for trivial equilibrium point since all the coefficients of polynomial $Q(\lambda)$ are positive, and hence, the trivial equilibrium point (T_1, I_1) is locally stable. For checking the stability of the nontrivial equilibrium point, the characteristic equation (A69) is evaluated in (T_2, I_2):

$$\lambda_{1,2}|_{(T_2,I_2)} : Q(\lambda)|_{(T_2,I_2)} = \lambda^2 + \underbrace{\left[eT_2 + f\right]}_{U}\lambda + \underbrace{\left[-\frac{nbh^nT_2}{I_2(h^n+I_2{}^n)}(d-eI_2)\right]}_{V} = 0 \tag{A70}$$

By assuming the condition (A67), the coefficients U and V are positive and negative respectively. Therefore, the sign of the coefficients of $Q(\lambda)$ (U and V) changes only once and hence, there exists an eigenvalue with positive real part. Therefore, the nontrivial equilibrium point (T_2, I_2) is a saddle node and unstable.

Similar to the model (1), the stable manifold of saddle node in the model (A63) defines a threshold for the initial conditions that allow for unlimited proliferation of activated T cells. By comparing the conditions (A67) and (3), the dependencies of these conditions to the model parameters, specifically the proliferation rate (a) and IL-2 secretion rate (d), are positively correlated. In other words, in both models, only T cell clones with sufficiently high proliferation rate (a) and/or high IL-2 secretion rate (d) are able to undergo major T cell proliferation.

For $k \neq 0$, the equilibrium points of the model (A63) are obtained from the following equations:

$$T = \frac{fI}{d-eI} \tag{A71}$$

$$I : \left[(a-b)f - ke\right]I^{n+1} + [kd]\,I^n - \left[h^n(bf+ke)\right]I + kdh^n = 0 \tag{A72}$$

The stability of equilibrium points is analyzed by evaluating the characteristic equation and is shown in **Figure A1** for parameter values given in **Table 1** and Hill-function parameters $n = 2$ and $h = 0.5$. By comparing the bifurcation diagram in **Figures A1** and **3B**, the qualitative similarity between model (4) and (A63) is evident. This qualitative similarity also holds true between model (13) and the following model:

$$\begin{cases} \dfrac{\mathrm{d}T}{\mathrm{d}t} = \Phi(I)T - bT - cT^2 - \gamma RT + k \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = \epsilon\Phi(I)R - bR + \beta\widehat{N} \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = dT - eI(T+R) - fI \end{cases} \tag{A73}$$

where $\Phi(I)$ is identical to (A64).

The equilibrium points of model (A73) are calculated and their stability is analyzed by deriving the characteristic equation of the model and obtaining the eigenvalues. By keeping assumption (12), the bifurcation diagrams of model (A73) for two different values of λ are obtained by treating $k = \frac{\beta N_0}{g+\beta} > 0$ as the bifurcation parameter (depicted in **Figure A2**). Depending on the value of k, the model has either 8 or 6 equilibrium points (4 or 2 equilibrium points with $T > 0$, identical to model (13)) with parameter values given in **Table 1** and Hill-function parameters $n = 4$ and $h = 1$. The additional equilibrium points resulted from considering the Hill-function nonlinearity are all in the negative space of the model variables. As it can be seen from **Figure A2**, similar to the model (13), the three qualitatively different responses still could be derived from the modified model. It is clear that the value of k_i, k_s, and λ_{th} are different from their corresponding values in the model (13).

In summary, imposing the nonlinear IL-2 dependent proliferation rate of cells results in a more restricted condition for initiation of an immune response in comparison to the linear IL-2 dependent proliferation rate, namely the requirement of higher T cell avidity (higher a and d), higher Ag-stimulation (increased β_i), and lower Treg/Tconv ratio (lower λ_{th}); but three qualitatively different immune reactions depending on the critical levels of Ag-stimulation could still be derived, very similar to model (13).

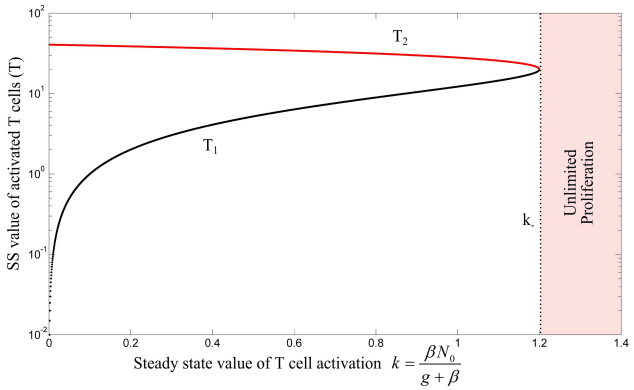


FIGURE A1 | Bifurcation diagram of model (A63) with Hill-function parameters $n = 2$ and $h = 0.5$ by treating k as bifurcation parameter. Stable and unstable equilibrium points are shown by black and red lines, respectively. For $k > k_-$, the immune response enters the regime of unlimited proliferation. The unstable negative equilibrium point is omitted in this figure.

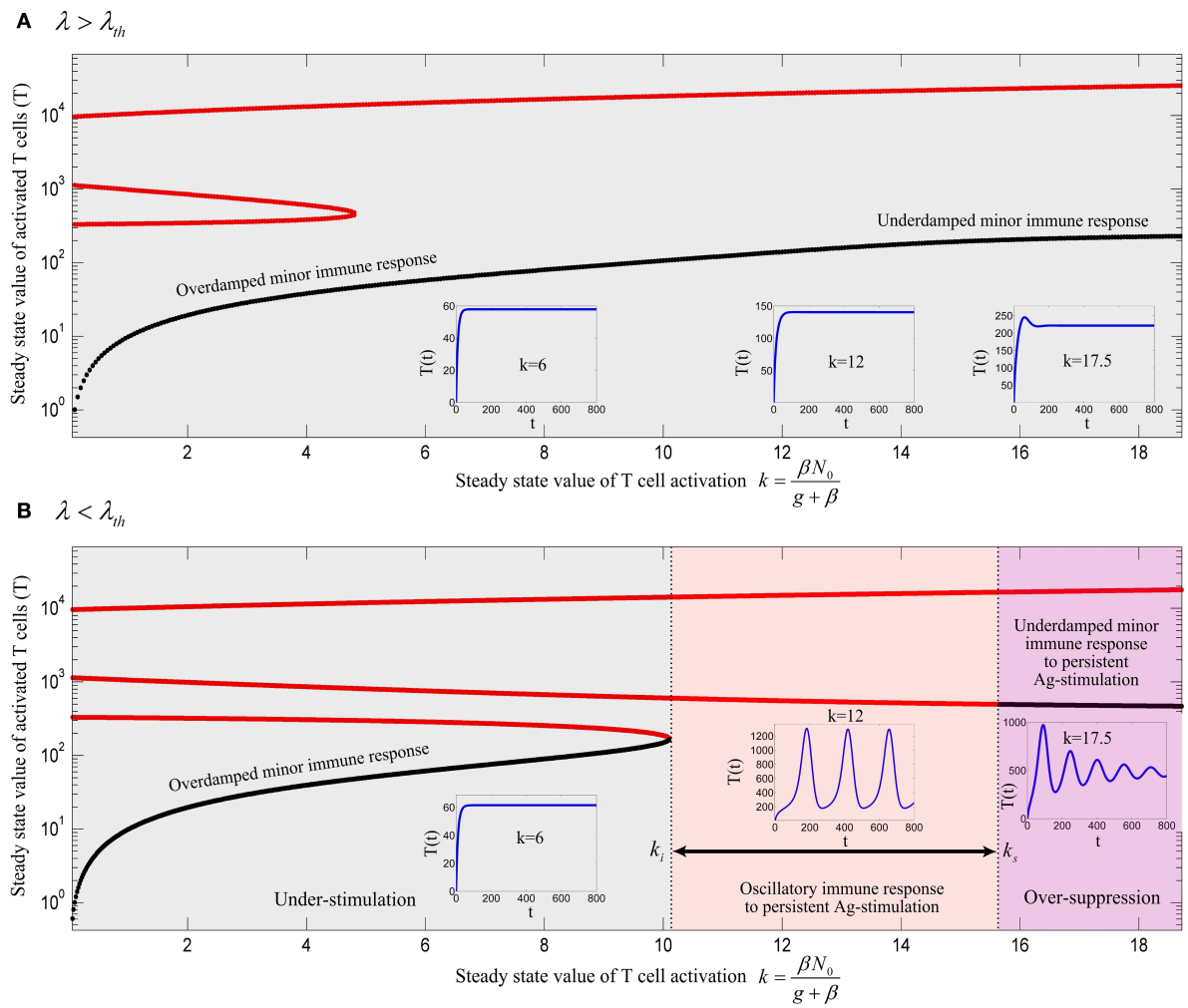


FIGURE A2 | Bifurcation diagram of model (A73) with Hill-function parameters $n = 4$ and $h = 1$ using k as the bifurcation parameter with (A) $\lambda = 0.0016$ and (B) $\lambda = 0.0008$. Stable and unstable equilibrium points are drawn by black and red solid lines, respectively. The stable limit cycles are not shown for all value of $k_i < k < k_s$ except for $k = 12$. Depending on the values of λ and k, an immune response is not initiated (gray), is initiated (red) or over-suppressed (magenta). With parameter values given in **Table 1** and Hill-function parameters $n = 4$ and $h = 1$, the threshold becomes $\lambda_{th} = 0.00111$. The time-courses of the activated T cell population $T(t)$ were deduced from a numerical solution of model (A73) with zero initial conditions and persistent β. The negative equilibrium points which are all unstable are not shown in the plots.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 February 2014
doi: 10.3389/fimmu.2014.00057

Harnessing the heterogeneity of T cell differentiation fate to fine-tune generation of effector and memory T cells

Chang Gong[1], Jennifer J. Linderman[2] and Denise Kirschner[3]*

[1] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
[2] *Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, USA*
[3] *Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA*

Edited by:
Veronika Zarnitsyna, Emory University, USA

Reviewed by:
Rustom Antia, Emory University, USA
Thomas Höfer, German Cancer Research Center, Germany

****Correspondence:***
Denise Kirschner, 6730 Medical Science Building II, Ann Arbor, MI, USA
e-mail: kirschne@umich.edu

Recent studies show that naïve T cells bearing identical T cell receptors experience heterogeneous differentiation and clonal expansion processes. The factors controlling this outcome are not well characterized, and their contributions to immune cell dynamics are similarly poorly understood. In this study, we develop a computational model to elaborate mechanisms occurring within and between two important physiological compartments, lymph nodes and blood, to determine how immune cell dynamics are controlled. Our multi-organ (multi-compartment) model integrates cellular and tissue level events and allows us to examine the heterogeneous differentiation of individual precursor cognate naïve T cells to generate both effector and memory T lymphocytes. Using this model, we simulate a hypothetical immune response and reproduce both primary and recall responses to infection. Increased numbers of antigen-bearing dendritic cells (DCs) are predicted to raise production of both effector and memory T cells, and distinct "sweet spots" of peptide-MHC levels on those DCs exist that favor CD4+ or CD8+ T cell differentiation toward either effector or memory cell phenotypes. This has important implications for vaccine development and immunotherapy.

Keywords: two-compartment model, lymph nodes, blood, agent-based, circulation, systems biology, dendritic cells, cognate

INTRODUCTION

Antigen-presenting cells (APCs), especially dendritic cells (DCs), process antigens and carry information from sites of infection to secondary lymphoid organs, such as lymph nodes (LNs) (1). T cells are produced in the thymus and are deployed into blood circulation to recognize millions of different epitopes from pathogenic organisms; each T cell is hardwired to have one type of T cell receptor (TCR) that recognizes a single pattern (i.e., "cognate" with respect to a specific antigen) (2). The frequency of particular cognate T cells is as low as 10^{-5}–10^{-6} (3, 4). Through high endothelial venules (HEVs), T cells are recruited to LNs, where they are exposed to antigenic peptides presented by MHC molecules expressed on DCs – this initiates the adaptive immune response (5–9). LNs are organized such that when T cells travel through they can be efficiently scanned by DCs to identify that rare cognate encounter (10–12). Such encounters result in binding of cognate T cells to DCs and subsequent activation and proliferation of the T cells. The expanded T cell population differentiates into two classes: effector cells, which perform immediate killing and cytokine secretion functions, and memory cells, which are reserved for long-term protection (13, 14). These cells move out of LNs via efferent lymphatics (ELs) into blood circulation (15). Through the blood, effector T cells reach sites of infection while memory T cells continue to recirculate and await a potential secondary infection for which they will wage a faster and stronger recall response (16, 17). A snapshot of the trafficking of these cells is shown in **Figure 1**. The immune system responds differently to different antigenic materials; however, the same set of machinery is engaged to face each challenge. Thus, there should be a general program adaptively guiding the behavior of this system. In this study, we focus on cellular-mediated events shared among immune responses during the initiation of adaptive immunity and generation of immune memory.

Differentiation of T cells during generation of adaptive and memory responses is highly heterogeneous, and this heterogeneity is may be dependent on the environmental context that each cell experiences (18, 19). However, the cause of such heterogeneity is poorly understood. If mechanisms other than mere stochasticity contribute to heterogeneity, it could be possible to more precisely direct the differentiation to favor the production of the desired output from an immune response (e.g., effectors in an immune therapy or memory cells in vaccination) by manipulating the mechanisms involved. We are interested in which mechanisms could provide handles for such manipulation. Since T cell priming occurs in LNs, and blood circulation conveys effector and memory T cells to locations where they perform their specific functions, mechanisms in these two organs could be responsible for the heterogeneous differentiation. The dynamics of T cells in these compartments will also reflect progression of infection or effectiveness of vaccinations. Thus, understanding how different LN and blood mechanisms affect the dynamics of infection and treatment could help guide immunotherapy and vaccine design.

Computational and mathematical models are widely used in biological systems to assess hypotheses and generate predictions

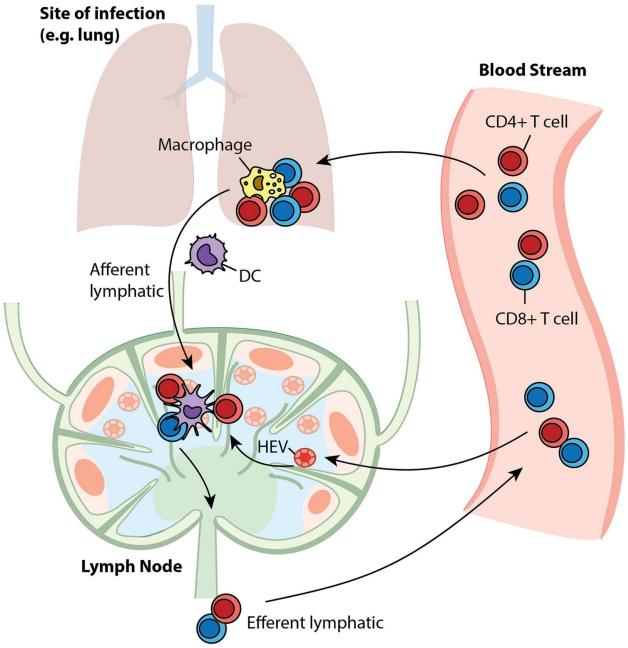


FIGURE 1 | T cell trafficking between compartments. Naïve T cells circulate between LNs and blood. Upon infection, APCs present antigen to cognate T cells in LNs to initiate their proliferation and differentiation to generate effector and memory cells. After entering the blood, effector cells are recruited to sites to fight ongoing infection, while memory cells recirculate, awaiting secondary infections.

for experimental validation. Deterministic equation-based models have been developed to understand the dynamics of T cells responding to immunogenic antigens, and these models helped with estimating parameters, determining alternative hypothesis, and predicting the outcomes of immune responses (20, 21). Agent-based models (ABMs) have proven convenient in assessing roles of cellular and molecular level interactions during infection (22–27). However, because of the extremely low cognate frequency that exists in primates, these models usually require large numbers of cells to be simulated and thus are very computationally intensive. In order to capture both heterogeneous stimuli-sensitive short-term activation events as well as average long-term dynamics, a model needs to be capable of adapting itself to both situations.

In this study, we present a hybrid computational model that uses an agent-based modeling to capture events occurring in a LN and a non-linear ordinary differential equation (ODE) model to capture events occurring in the well-mixed compartment of blood. This model allows us to track a highly stochastic immune response operating during the first few weeks of an immune response (with time resolution around seconds), as well as long-term dynamics afterward (at a time scale of months to years). Using this model, we assess which mechanisms in both LN and blood compartments control the differentiation and clonal expansion processes of T cells and also direct the immune response toward potent effector T cell output and/or robust memory generation. These findings could bring insights to vaccine design strategies.

MATERIALS AND METHODS

LN ABM MODEL

Agent-based models are computational models in which individual agents are represented on an explicitly formulated grid and they interact with each other according to a defined set of rules implemented in discrete time steps. As these types of models can account for spatial-sensitive interactions between DC and cognate T cells, they are ideal for studying heterogeneous priming and differentiation of T cells in LNs (23–25, 28, 29).

We previously developed *LymphSim*, a three-dimensional (3D) LN computational model capturing dynamics of CD4+ T cells, CD8+ T cells, and DCs during both steady state and infection (30). Briefly, cells move on a 3D grid that is shaped like a truncated cone and represents ~1/200 of a primate LN. T cells enter the LN via HEVs, search for DCs, activate and proliferate to generate effector cells that exit via ELs. In *LymphSim*, cell motility and steady state values in a LN are calibrated to experimental data with model antigens such as OVA (31), and the dynamics during an immune response are not quantitatively fit to any specific infection. For simplicity, we only include one type of cognate T cell each for CD4+ and CD8+ T cells in current model, and DCs present the corresponding antigens on pMHC (peptide-MHC)-II and pMHC-I for both primary and secondary infections. The model can be adapted to account for multiple sub-antigens. For the work herein, this single antigen study is sufficient to address the key questions under study. A complete list of rules can be found at: http://malthus.micro.med.umich.edu/lab/movies/3dLN/.

EFFECTOR AND MEMORY T CELL DIFFERENTIATION RULES

In the present study, we modified *LymphSim* to include two additional T cell differentiation states: central memory (CM) and effector memory (EM), for both CD4+ and CD8+ T cells. We also added rules that govern generation of these memory cells, and their interaction with other cells (**Figure 2**).

We based the cell differentiation process on a version of a "signal-strength model," in which the overall strength of signal received by a naïve T cell during DC contact will determine the fate of cell differentiation (**Figure 3**) (32–35). A definitive differentiation scheme after T cell priming occurs has not been determined by experimentation. Previous modeling studies based on experimental data reject memory to effector differentiation in favor of effector to memory differentiation (20); however, more recent work showed that differentiation has as its backbone differentiation from naïve to CM precursor to EM precursor to effector (18). The scheme we use in this study considers effector to EM differentiation, but is still topologically similar to the scheme from (18), with precursors of both EM and effectors differentiating into these two subtypes (**Figure 3**). The difference between the two schemes is that "effectors" in our model are cells that have differentiated toward effector phenotype sufficiently so as not to enter into the CM population, nor have they entered into the EM pool. They are allowed to exit the LN due to the loss of early activation markers (CD69), even though these cells do not perform effector functions until they would reach sites of infection, which is not studied in this current work.

In our model, a series of probabilistic checkpoints are established to determine to which state a cell will proceed (36–39).

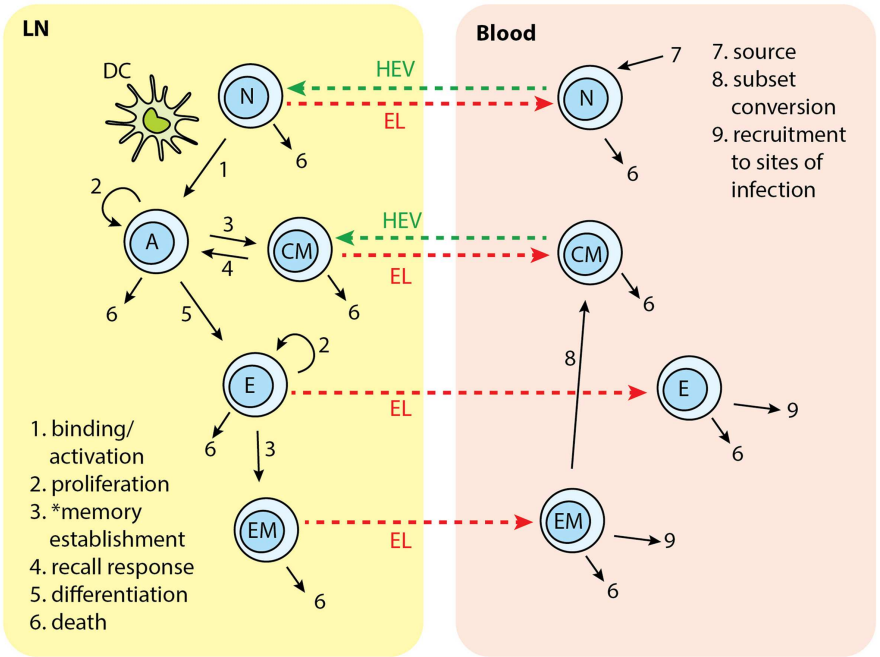


FIGURE 2 | T cell subsets in two-compartments of LNs and blood: N, naïve; A, activated; CM, central memory; E, effector; EM, effector memory. Each number indicates a collection of processes occurring in that step and in different cell types. Naïve T cells are recruited to LN from blood. In the LN, cognate T cells bind with Ag-DCs and get activated. Activated T cells proliferate and differentiate into central memory (CM) and effector cells. CM in the LN can bind to DC and be activated again. Effector T cells can further differentiate to effector memory (EM) cells. Naïve, effector, CM, and EM exit LN from EL. Naïve and CM cells recirculate between LN and blood. Effector and EM are recruited to sites of infection. EM can covert to CMs. *Memory establishment for CD8+ T cells requires LDCs.

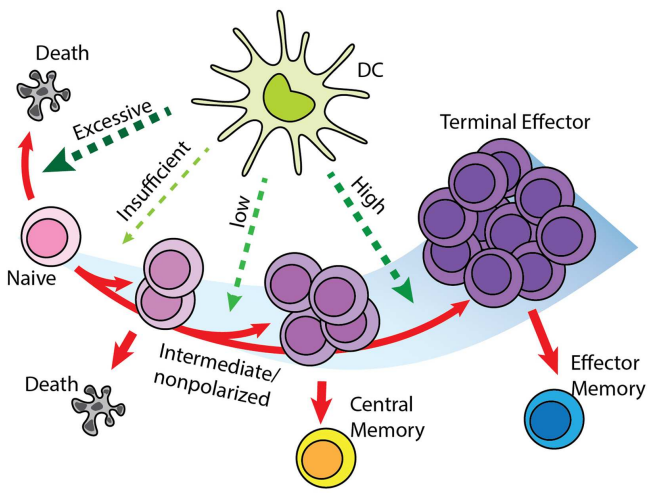


FIGURE 3 | "Signal-strength model" of T cell differentiation. T cells receive antigenic, co-stimulatory, and inflammatory signals from DC during priming. In concert, these of stimulations determine the fate of T cell clonal expansion and differentiation. Greater proliferation correlates with stronger signal. However, insufficient stimulation results in death by neglect, while excessive stimulation causes activation induced cell death. Stronger stimulation also drives T cells toward terminal differentiation and reduces their memory-forming potential. Please see Section "Effector and Memory T Cell Differentiation Rules" for a description of differentiation models and how this was selected.

When a cognate T cell finds an Ag-bearing DC (Ag-DC) or licensed DC (LDC) in its binding area, the corresponding pMHC value of the DC is checked to see if a successful binding can be established. If bound, a T cell continuously accumulates signals from the DC (40), represented by pMHC levels at each time point. Here, pMHC level is used as a proxy for the strength of antigenic stimulation from the DC or LDC. When a T cell unbinds from a DC or LDC, the accumulated signal value is used to determine whether a T cell proceeds to an activated state, or returns to a resting state (naïve). Activated cells go through a set number of rounds of divisions, after which the accumulated signal level is checked again to decide if the cell can further differentiate into an effector state. Effector cells will divide a few more rounds. With given probabilities, the cells with intermediate differentiation status do not proceed to effector status, but become CM cells, while those effector cells with sufficient signals will become EM cells (41–43). The probability of effector cell converting to EM is estimated between 0.1 and 0.4. CM T cells can be recruited to LNs from HEVs. These cells act similarly to cognate naïve T cells. When they detect Ag-DCs or LDCs, CMs will bind to DC and accumulate signal more efficiently in comparison with naïve cells (44, 45). The rules above apply to both CD4+ and CD8+ T cells. Because we developed some of these rules based on LCMV studies, one difference we captured between these two cell types is that CD8+ T cells can bind only to LDCs to generate functional memory cells in the primary response, whereas CD4+ T cells do not have this restriction and can generate memory cells after binding to both Ag-DCs and LDCs (46).

Other models of T cell differentiation exist, and some of these models are not mutually exclusive. We also integrated features from these models into our rule set, and excluded those that are inconsistent with current findings or are not applicable to our model at this stage. A single naïve T cell can produce both effector and memory progenies (19, 47), so we excluded the possibility that effector and memory arise from separate precursors. In the decreasing-potential model (13), the stimulation that T cells receive during infection drives greater clonal expansion but reduces their potential to differentiate into memory cells. Some studies show T cells are committed to massive proliferation after initial encounters with APCs, and can differentiate into both memory and effector subsets even if adoptively transferred into hosts absent of antigen (48). Thus, we limited the signal accumulation stage to the period of time when a T cell is bound to a DC before its first division, similar to findings made in B cell expansion (49). In the asymmetric cell fate model (50), heterogeneity arises from the unequal distribution of differentiation factors into daughter cells during division. We will further study this hypothesis as we incorporate dynamics at molecular level, but currently account for these asymmetries using phenomenological probabilities.

BLOOD COMPARTMENT SUB-MODEL: ODE AND PARAMETER ESTIMATION

We developed a blood compartment model by assuming the blood is a well-mixed, homogenous compartment. We use a system of non-linear ODEs to capture the dynamics of T cells therein. Equations for CD4+ T cells are:

$$\frac{dN_4}{dt} = S_{N4}(t) - \delta_{N4}N_4 + e_{N4}^{\mathrm{LN}} \quad (1)$$

$$\frac{dE_4}{dt} = -\delta_{E4}E_4 - \xi_{E4}E_4 + e_{E4}^{\mathrm{LN}} \quad (2)$$

$$\frac{d\mathrm{CM}_4}{dt} = -\delta_{\mathrm{CM4}}\mathrm{CM}_4 + \alpha_{\mathrm{EM4}}\mathrm{EM}_4 + e_{\mathrm{EM4}}^{\mathrm{LN}} \quad (3)$$

$$\frac{d\mathrm{EM}_4}{dt} = -\delta_{\mathrm{EM4}}\mathrm{EM}_4 - \xi_{\mathrm{EM4}}\mathrm{EM}_4 - \alpha_{\mathrm{EM4}}\mathrm{EM}_4 + e_{\mathrm{EM4}}^{\mathrm{LN}} \quad (4)$$

N_4, E_4, CM_4, and EM_4 represent the blood concentrations of naïve, effector, CM, and EM CD4+ T cells, respectively. $S_{N4}(t)$ is the time-dependent thymus output of naïve CD4+ T cells (51). The initial output is estimated from healthy 30-year-old individuals, and declines by 5% per year (52). δ_{N4} is the overall death rate constant for naïve cells, including homeostatic proliferation and death. We estimated this parameter by assuming a quasi-equilibrium between thymus output and peripheral loss. δ_{E4}, δ_{CM4}, and δ_{EM4} are the death rate constants for effector, CM, and EM CD4+ T cells, respectively. δ_{E4} and δ_{EM4} account for the death of circulating effector and EM cells, excluding those recruited to sites of infection (53). δ_{CM4} reflects the overall loss of CM cells, including self-renewal and death (53). ξ_{E4} and ξ_{EM4} are the rate constants for recruitment of CD4+ effectors and EM cells from blood to sites of infection. As the dynamics at a site of infection are not considered in this study, these recruitment terms serve as a sink for the corresponding cell species in the blood compartment. α_{EM4} is the rate constant for EM cell differentiation into CM cells (54). The terms e_{N4}^{LN}, e_{E4}^{LN}, $e_{\mathrm{CM4}}^{\mathrm{LN}}$,and $e_{\mathrm{EM4}}^{\mathrm{LN}}$ represent rates of LN net output of corresponding cells. These terms are converted to the changes in concentration in the blood per time step. For naïve and CM cells, this is calculated as the difference between the number of exited and recruited cells. For effector and EM cells, this is calculated as the number of exited cells. These four terms are not solved directly in the ODE system but rather are added as an initial condition before each blood time step is processed in the computational model. We show them in the equations for completeness. Similar equations and parameter estimates are written for CD8+ T cells (see Supplementary Material). Because the CM CD8+ T cells population is maintained for life, we assume a very small value for the loss rate constant δ_{CM8}, corresponding to half-life of 20 years (53). See Table S3 in Supplementary Material for a complete list of parameters, definitions, values, units, and source references.

TWO-COMPARTMENT HYBRID MODEL

Our goal is to develop a two-compartment computational model that combines *LymphSim* and the blood ODE model described above. Recently, we published other models linking ODEs and ABMs (55–57). For this study, we use the implementation method we employed successfully to link a LN compartment with a lung (56). The LN and blood compartment models are processed sequentially during each time step of simulation (**Figure 4**). During the T cell recruitment subroutine of the LN ABM model, the probability of recruiting T cells of each type/state is calculated based on their blood concentration levels. At the end of LN compartment simulation time step, the LN net output is calculated as the difference between exited and recruited number of each cell type and is multiplied by a factor that accounts for physiological compartment-size scaling from 0.5% back to the entire paracortex and unit conversion from cell number to blood concentration. This net output is then added to the corresponding variables in blood compartment ODEs. We have made a few assumptions regarding how we capture the LN to blood dynamics. First, we are only modeling dynamics of T cells and DCs within a single LN. There are ~700 LNs in the human body and they are connected via an intricate lymphoreticular network. T cells travel between multiple LNs via these lymphatics and eventually enter the blood via the superior vena cava. We assume that cells exit the LN and enter the blood compartment immediately, coarse-graining the time spent in the lymphatic system. However, our cells travel through the LN and blood in time frames consistent with experimental data [<24 h; Ref. (58)], accounting for the delay.

For computational efficiency, we use a method we term *tunable resolution* (TR) (manuscript submitted). One of the goals of TR is to develop multi-scale models with sub-models of different resolutions, so that models can be run with coarse- or fine-grained alternative versions of sub-models during simulation to save resources without sacrificing accuracy. Here, for each physiological compartment (blood or LN), there is a computational switch that allows the model in an automated fashion to bypass simulation of a given compartment. In this two-compartment model, we do not have an alternate version of each compartment *per se*; instead, each compartment can be suspended when specific criteria are met. For example, during the pre-simulation,

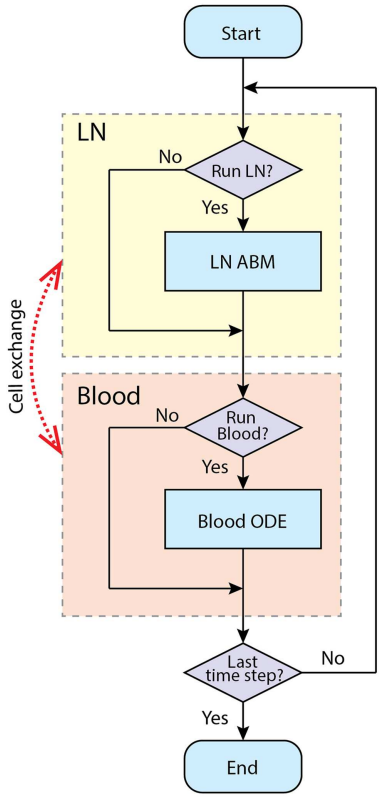


FIGURE 4 | Simulation procedure using tunable resolution. LN and blood compartment are processed sequentially in each time step. LNs recruit cells from blood and put exiting cells into it. The recruitment probabilities are modified by blood concentration of corresponding cell types. Each compartment has a switch to determine whether this compartment is processed in current time step, or will it be bypassed.

the blood compartment is turned off, and the LN is simulated until a baseline steady state is reached. When an immune response is occurring, LN and blood compartments are both running to simulate the immune response in fine-grained, spatially explicit detail for a time scale of a few weeks. When an active immune response finishes and there are no Ag-DCs, LDCs, bound, active, effector, or EM cells in the LN compartment, the LN compartment is suspended to allow rapid simulation of the blood compartment at longer time scales (months to years). When a secondary infection begins, the LN compartment is switched on again (**Figure 4**).

MODEL CALIBRATION

The hybrid model contains 103 parameters that govern mechanisms occurring in both physiological compartments and the interactions between them (see Table S3 in Supplementary Material for a complete list of the parameters). For the LN compartment model, parameters governing T cell motility and trafficking are calibrated to data as described previously (30). Parameter estimates for the ODE model in the blood compartment are discussed in Section "Blood Compartment Sub-Model: ODE and Parameter Estimation" and Supplementary Material.

To use our model for memory T cell differentiation dynamics, we estimated parameters in our model using the limited data available in the literature for memory cell generation in LNs. We estimated parameters governing total production of expanded cognate CD8+ cells generated in the LN model (Table S3 in Supplementary Material, parameters marked with ‡) to data from T cell clonal expansion studies in mice using OVA as a stimulating antigen (59). In that study, DCs are ablated at different time points to show that the duration of antigen presentation correlates with magnitude of T cell expansion, but a short exposure is sufficient to program CD8+ T cells to differentiate into both effector and memory subsets (59). We adapted our model to reflect experimental methods used in these studies. DCs are removed from the LN grid at indicated time points after the recruitment during primary challenge (**Figure 5A**), as was done experimentally by injecting diptheria toxin (DT) (59). Unlike rules for LCMV as previously discussed, CD8+ naïve T cells are allowed to bind both Ag-DC and LDC to be primed and the enter memory state. This is because in these experiments, DCs are activated from LPS pulsing or *Listeria monocytogenes*-OVA. From our *in silico* experiments terminating antigen presentation from DCs at various time points after Ag-DC recruitment, we predict that the magnitude of the primary response is dependent of the duration of DC presence (**Figure 5B**). However, a very short period of stimulation is capable of generating memory cells, as we see a potent production of antigen-specific CD8+ T cells after a secondary challenge (**Figure 5C**). Moreover, it takes only 3 days for the recall response to exceed the magnitude of primary response on day 5, indicating a faster reaction to previously experienced antigens, as observed *in vivo*. Our simulation results are comparable to data from the Prlic study (59). The parameter set we obtained is used as our baseline for simulating infection scenarios (see below and Table S3 in Supplementary Material).

SIMULATED INFECTION AND MODEL VALIDATION

We next validated our model with data sets from experimental studies using LCMV or OVA as stimulating antigens. For each simulated infection, a 3-day pre-simulation of the ABM LN sub-model precedes the actual experiment to allow cells to reach a steady state in terms of quantity and spatial distribution. During this period, the blood sub-model is suspended, with the naïve CD4+ and CD8+ T cells concentrations fixed at 450 and 320/mm^3, respectively (51, 60). Then Ag-DCs are introduced to stimulate the T cell response. We represent this by introducing antigen-bearing DCs in such a way as to mimic an acute infection (23, 25, 30). Ag-DCs carry and present a unique antigen and are recruited to the LN compartment for 2 days. These DCs will prime cognate T cells (cognate frequency is set to 10^{-4}) for about 5 days before they die, mimicking a hypothetical acute infection. To mimic a hypothetical secondary infection, Ag-DCs are recruited to the LN again from day 600 to 602. Each experiment is simulated five times to reduce aleatory uncertainty.

To confirm that our model produces reasonable dynamics in the blood compartment, we qualitatively compare the time course of blood antigen-specific cells to data sets from LCMV studies, where the measurements are performed in spleen (61–63).

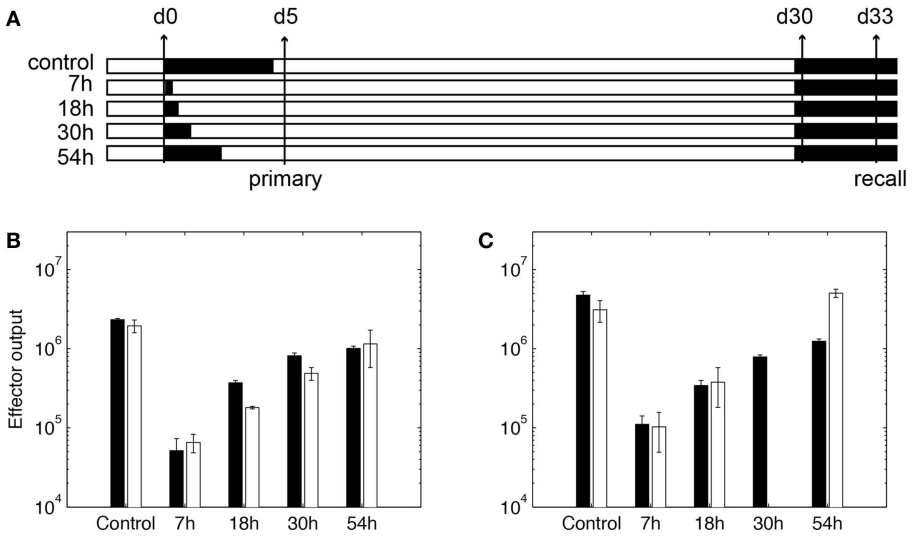


FIGURE 5 | Expansion of CD8+ T cells in simulation. **(A)** *In silico* experimental schemes. Black bars show the duration of Ag-DC presence. In primary challenge, DC antigen presentation is terminated at time points indicated on the left. "Control" indicates no termination and DC are allowed to live their natural lifespan. Recall challenge is given from day 30. Measurements are taken on day 5 for primary response and day 33 for recall response, respectively. **(B,C)** CD8+ T cell population in simulated responses (black bars) and experimental data (white bars) (59). **(B)** Size of expanded CD8+ T cell population in primary response. Values are measured from day 5 after Ag-DC are recruited to the LN. **(C)** Size of expanded CD8+ T cell population during recall response on day 33. *X*-axis value indicates antigen presentation times in the primary challenge.

For lineage tracing simulations (see Cognate Naïve Cells Undergo Heterogeneous Expansion), every cognate naïve T cell recruited to the LN is assigned a unique serial number. This number is passed to the daughter cells when these labeled T cells proliferate and differentiate, so T cells sharing the same serial number belong to the same single-cell derived progeny. When each differentiated cell exits the LN, its serial number is recorded. For each individual cognate precursor, the number of descendant cells and their differentiation states are calculated. Cell progenies are ranked by the abundance of their progenies, from largest to smallest. Five replications are performed for this simulation.

We calculate the Index of disparity D between the expanded populations of different single-cell derived progenies (19), which is the inverse Simpson diversity index mapped to a 0–1 interval:

$$D = \frac{N - \frac{1}{\sum_{i=1}^{N} f_i^2}}{N - 1}, \tag{5}$$

where N is the number of progenies and f is the frequency of each single-cell derived progeny in the total population.

UNCERTAINTY AND SENSITIVITY ANALYSIS

In this study, our goal is to reproduce patterns of generalized immune responses. In addition to using parameters estimated from previous work (see above and Table S3 in Supplementary Material) and experimental data, we use global uncertainty and sensitivity analysis (U/SA) to study how particular biological mechanisms affect simulation outputs (64).

For each set of sensitivity analysis, a list of parameters is chosen, and for each parameter of interest, a range is specified. Latin hypercube sampling (LHS) is applied to generate the matrix of parameter values, where each experiment represents one combination of sampled parameter values. LHS is a stratified sampling method that requires fewer samples compared with random sampling method but achieves the same accuracy (65). This technique is particularly helpful for our ABM model where parameter values need to be estimated from a high-dimensional space (64). The parameter space is sampled completely and accurately, with a large sampling size. Each experiment is replicated five times to reduce aleatory uncertainty from inherent stochastic variations (64). After the simulation, model readouts are chosen and partial rank correlation coefficients (PRCCs) are calculated between each readout-parameter pair to assess global sensitivity and detect monotonic relationship between mechanisms and output of interests.

To study how various mechanisms affect the generation of memory from within each compartment (blood and LN) as well as how they influence the other compartment (LN and blood, respectively), we performed intra- and inter-compartment sensitivity analysis (64). We choose two sets of parameters governing mechanisms in each sub-model and estimated a range for each parameter (see Table S3 in Supplementary Material). We use LHS to sample the parameter space and generate 100 or 408 experiments for blood and LN experiments, respectively. Here we performed 2540 simulations, which provides ample coverage of the space. Sensitivities of outputs to mechanisms are assessed with PRCCs.

COMPUTATIONAL SIMULATIONS AND IMPLEMENTATION

Our hybrid model is implemented in C++ and runs on Linux/Mac OS/windows. Documentation and pseudo code are available in the online Supplement. A Forward Euler method is used to solve the ODEs. Each time step of the ABM simulation is further divided into 100 pieces (step size of 0.25 s) to reduce error. Each simulation

of 350 days (LN sub-model is active for ~40 days) takes 30–40 h to run.

RESULTS

HEALTHY UNINFECTED BASELINE DYNAMICS OF T CELLS ARE REACHED WITHOUT SIMULATED ANTIGEN PRESENTATION

Without Ag-DC introduced to the LN, all T cells remain naïve. The cell dynamics in LN in the absence of any infection present show that over a short time scale (days to weeks), cells remain at the steady state of ~170,000, or 4.0×10^6 cells/mm^3. There is an equal input/output flow of ~1000 cells per million cells per minute, and an average transit time of 16 h, which is consistent with previous data (30). The population of both CD4+ and CD8+ T cells in the blood declines long-term (20 years). By the end of year 20, the blood concentration of naïve CD4+ T cells drops to 210 mm^{-3}, and that of naïve CD8+ T cell drops to 170 mm^{-3}. Such long-term decline of naïve T cell number is comparable to clinical observations (66, 67).

EFFECTOR AND MEMORY T CELL POPULATIONS ARE GENERATED IN A SIMULATED ACUTE INFECTION

We simulated immune responses to a hypothetical acute infection by introducing Ag-DCs into the LN compartment to activate cognate T cells as shown in **Figure 6A**. The cognate frequency is set to 10^{-4}. **Figures 6B–E** show simulated immune cell dynamics in the LN and blood compartments.

In the LN compartment, the Ag-DC population increases first. These DCs scan surrounding CD4+ and CD8+ T cells and bind to their cognate matches. After this binding event, CD4+ and CD8+ T cells begin to proliferate and differentiate into active and effector T cells. After day 2, the influx of Ag-DCs to LN stops, and the number of Ag-DCs begins to decline (**Figure 6A**). At the same time, differentiated effector CD4+ T cells license Ag-DCs, further increasing their surface pMHC levels and stimulation strength, enabling them to allow CD8+ T cell memory potential. Because we assume that CD8+ T cell memory establishment requires LDCs, the appearance of CM and EM CD8+ T cells is delayed as compared with corresponding CD4+ T cells. After differentiation from the active state, effector, CM, and EM cells can exit the LN from ELs, resulting in the decline of these populations within the LN. The system eventually returns to baseline, but CMs can still recirculate through LN (**Figures 6B,C**).

In the blood compartment, the concentrations of effector, CM, and EM cell populations increase as they exit the LN (**Figures 6D,E**). The total concentration of both CD4+ and CD8+ Ag-specific T cell (effector, EM, and CM) peaks at about day 6 and 8, respectively (0.91 mm^{-3} for CD4+ T cells and 2.49 mm^{-3} for CD8+ T cells). The lifespans of effector cells are relatively short. These cells either die, or are recruited to sites of infection, bringing about a contraction phase characterized by a decline of total blood Ag-specific T cells. However, about 5% of the peak level is maintained in the memory cell class, especially CM cells in the

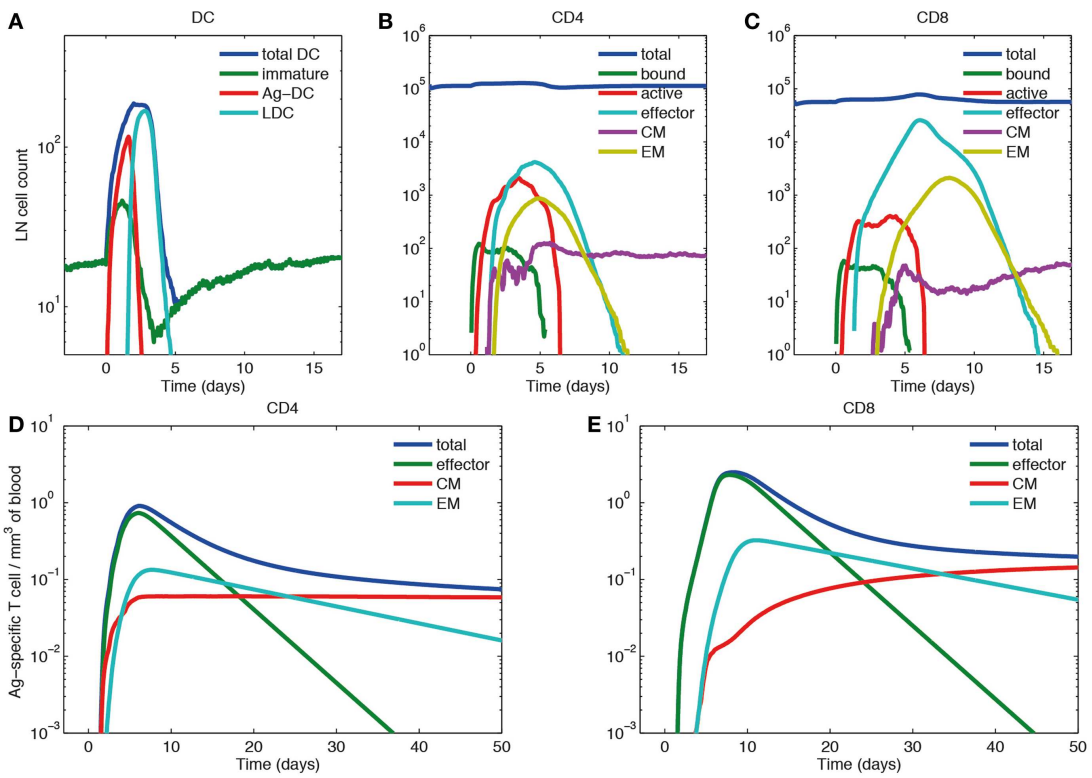


FIGURE 6 | Primary response dynamics of immune cells in LN and blood during a hypothetical acute infection (log scale). **(A–C)** Number of dendritic cells, CD4+ and CD8+ T cells of different subsets in the LN compartment. **(D,E)** Concentration of CD4+ and CD8+ T cells of different subsets in the blood compartment. **(A)** Model input of DCs representing a hypothetical acute infection, such as LCMV and **(B–E)** output.

long-term as their lifespan is longer than EM cells. In the blood, some of the EM cells convert to CM cells, while others are recruited away to sites of infection (**Figures 6D,E**). While there are no data from primates on these dynamics, our results are qualitatively in accordance with experimental data from mouse LCMV studies (61–63).

IMMUNE CELLS REACH HIGHER LEVELS DURING A RECALL RESPONSE AS COMPARED TO A PRIMARY RESPONSE

To understand the dynamics of a recall response, we simulated a scenario where Ag-DCs are introduced from day 0 to 2 in an initial round of infection (the same as that of Section "Effector and Memory T Cell Populations are Generated in a Simulated Acute Infection"). Once that infection dampens and immune cells return to a resting state, we introduce a second round of challenge by recruiting Ag-DCs from day 600 to 602. We challenge with the same antigen and use the same cognate frequency for naïve cells, but CM populations are maintained in the blood after the primary response. The resulting dynamics of Ag-specific T cells occurring in blood are shown in **Figure 6**.

As above, the primary response is initiated after the first round of Ag-DC input. Blood Ag-specific T cell numbers rise as the response continues and peak at day 6 and 8. After the peak, effector and EM T cells decline while the CM cell population is maintained. On day 600, the blood concentration of CM CD4+ T cells has dropped from 0.059 to 0.023 mm^{-3}, while the CM CD8+ T cell population remains at 0.16 mm^{-3}. The stable maintenance of CD8+ memory and decline of CD4+ memory is in agreement with mouse LCMV infection data (53). During the recall response, because of a memory cell population generated during the primary response that can faster and more strongly respond to the same antigen, both CD4+ and CD8+ T cells in the blood exceed peak levels of their primary response, peaking at 1.07 mm^{-3} for CD4+ and 6.05 mm^{-3} for CD8+ T cells. The recall response is more than twice as large as primary response for CD8+ T cells, but only marginally increased (18%) for CD4+ T cells. Such differences in CD4+ and CD8+ recall responses have been observed in LCMV experiments as well (68). After the recall response, higher levels of CM cells are maintained as compared to following the primary response (**Figure 7**). After the recall response ceases, the

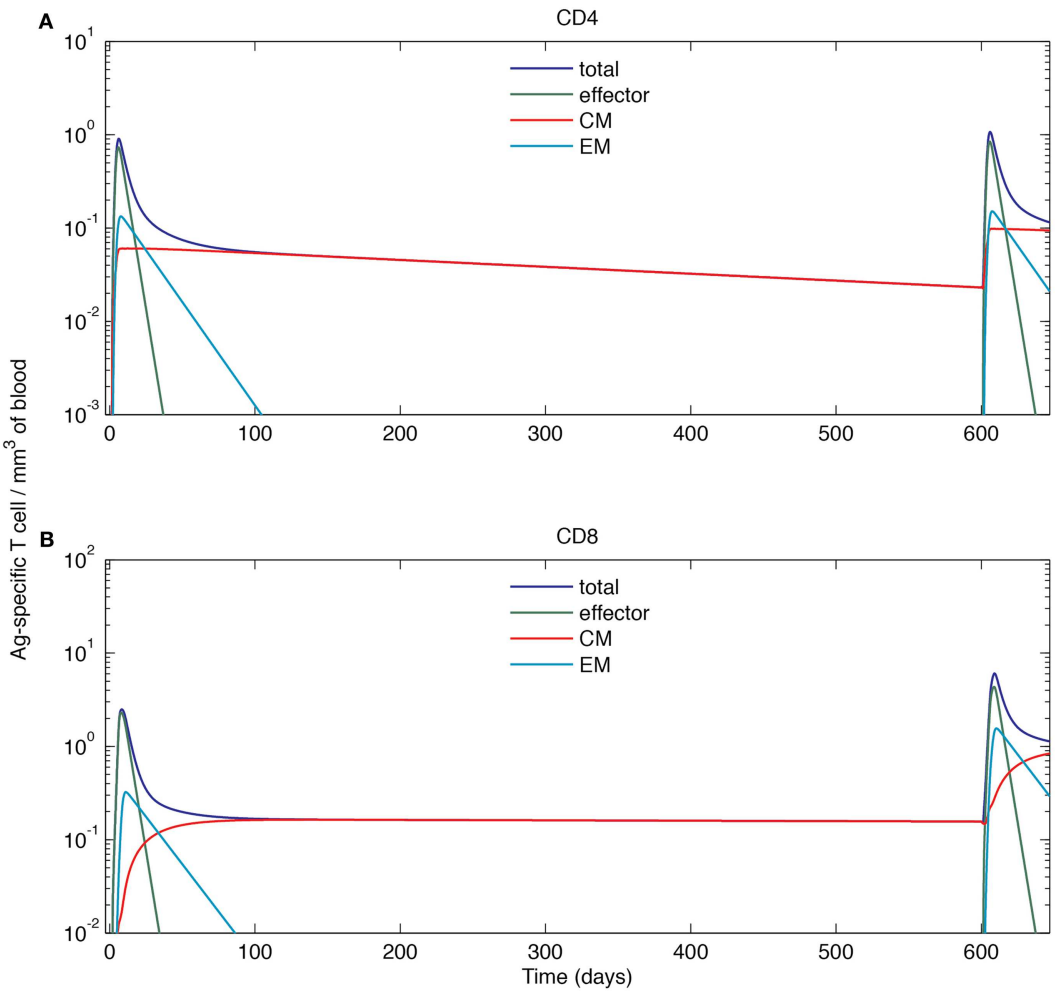


FIGURE 7 | Simulated cell dynamics in the blood compartment during a primary and recall response to a hypothetical acute infection (log scale). **(A)** Concentration of CD4+ T cells of different subsets in the blood. **(B)** Concentration of CD8+ T cells of different subsets in the blood. The left parts of the graphs are identical to those of **Figure 6**.

blood concentrations of CM cells are 0.094 and 0.84 mm^{-3} for CD4+ and CD8+ T cells, respectively. These results indicate that the antigen-specific immune memory is reinforced after the second round of challenge, as the central memory population formed in the primary challenge gets further expanded during the second round of challenge.

COGNATE NAÏVE CELLS UNDERGO HETEROGENEOUS EXPANSION

Recent lineage tracing studies showed that CD8+ T cells have a heterogeneous differentiation pattern (18, 19). Because the LN compartment of our model is agent-based, it is possible to track the fate of each individual cell during a simulated infection. We take advantage of this feature to validate our model using data from these recent studies.

Figure 8 shows our lineage tracing analysis for the primary response. The fraction of each single-cell derived progeny in the total population is shown in **Figure 8A** for CD8+ and **Figure 8E** for CD4+ T cells. For both CD4+ and CD8+ T cells, a small number of progenies have a large expanded population. The average size of the largest population is ~2000 for CD4+ T cells and ~8000 for CD8+ T cells. However, the majority of derived progenies have intermediate to small population sizes, with about 50 for CD4+ T cells and 200 for CD8+ T cells. The maximum population size of CD8+ T cell is 50-fold larger than the median. The index of disparity in our simulations is 0.81 for CD8+ T cells, close to the range of 0.85–0.95 shown in Ref. (19). These results indicate our model matches well with the heterogeneous differentiation experimental observations. While the corresponding experimental studies were performed only for CD8+ T cells, we are able to use our model to simulate the dynamics of CD4+ T cells as well. Our model predicts less heterogeneity for CD4+ T cells, with an index of disparity of 0.82 and a 50-fold difference between largest and median progenies (**Figure 8F**).

We also assessed the composition of these sub-populations. In **Figures 8C,G**, the proportion of CM cells of each single-cell derived progeny is plotted against the population size. These results suggest that progenies with a higher proportion of CM cells tend to have a smaller expanded population during the primary response. We calculated the Spearman correlation coefficients between each subtype and the total number of expanded cells (**Figures 8D,H**). The correlations are strong between effectors and overall total population size, but weak between CM and the overall population size. This is comparable to the results from Ref. (18). Thus, in addition to our other findings, these results confirm those previously identified (18, 19) that the magnitude of the primary response for single-cell derived progenies might not be the sole predictor of immune memory. We next study, which mechanisms influence the heterogeneous differentiation and clonal expansion processes of T cells.

ANTIGEN PRESENTATION BY DCs INFLUENCES OUTCOME OF AN IMMUNE RESPONSE

It is no surprise that antigen stimulation plays a crucial role in T cell activation and differentiation (35, 69). However, it is difficult to conduct experiments that quantitatively determine the mechanism of dependency. Here we varied the number of Ag-DCs recruited to LN in a range of 50–300 and the levels of pMHC molecules on the surface Ag-DCs from 100 to 300, and analyzed how they influence production of effector and CM cells. Model pMHC levels are used as a proxy for DC stimulation strength. Results are shown in **Figure 9**.

Increasing the number of Ag-DC recruitment promotes the output of both effector and memory T cells from the LN. The number of Ag-DC has a larger impact when the pMHC molecule levels are low. This result indicates that more Ag-DCs are beneficial for the production of higher levels of effector and memory

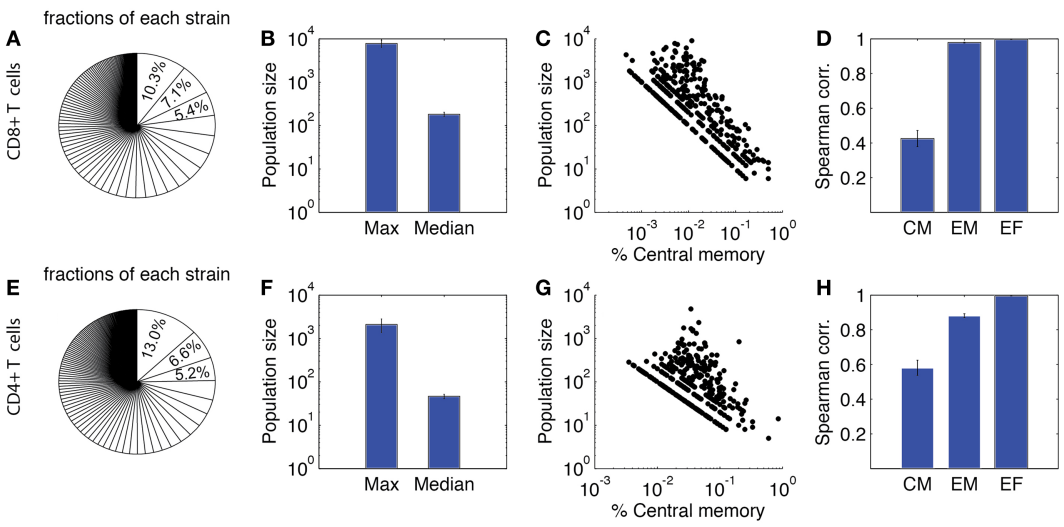


FIGURE 8 | Heterogeneity of expanded T cell families. Upper panel: CD8+ T cells. **(A)** Sizes of expanded CD4+ T cell population from each single-cell derived progeny (strain), represented as fraction of total expanded population as done in Buchholz et al. (18) and Gerlach et al. (19). **(B)** Maximum and median size of CD4+ T cell progenies. **(C)** Correlation between population size and the percentage of central memory cells. Each dot represents progeny of a single-cell derived progeny. **(D)** Spearman correlation of CM, EM, and effector CD4+ T cell number versus total expanded population. **(E–H)** Prediction of CD4+ T cell heterogeneity in clonal expansion. The four panels correspond to the same readouts as for panels **A–D**.

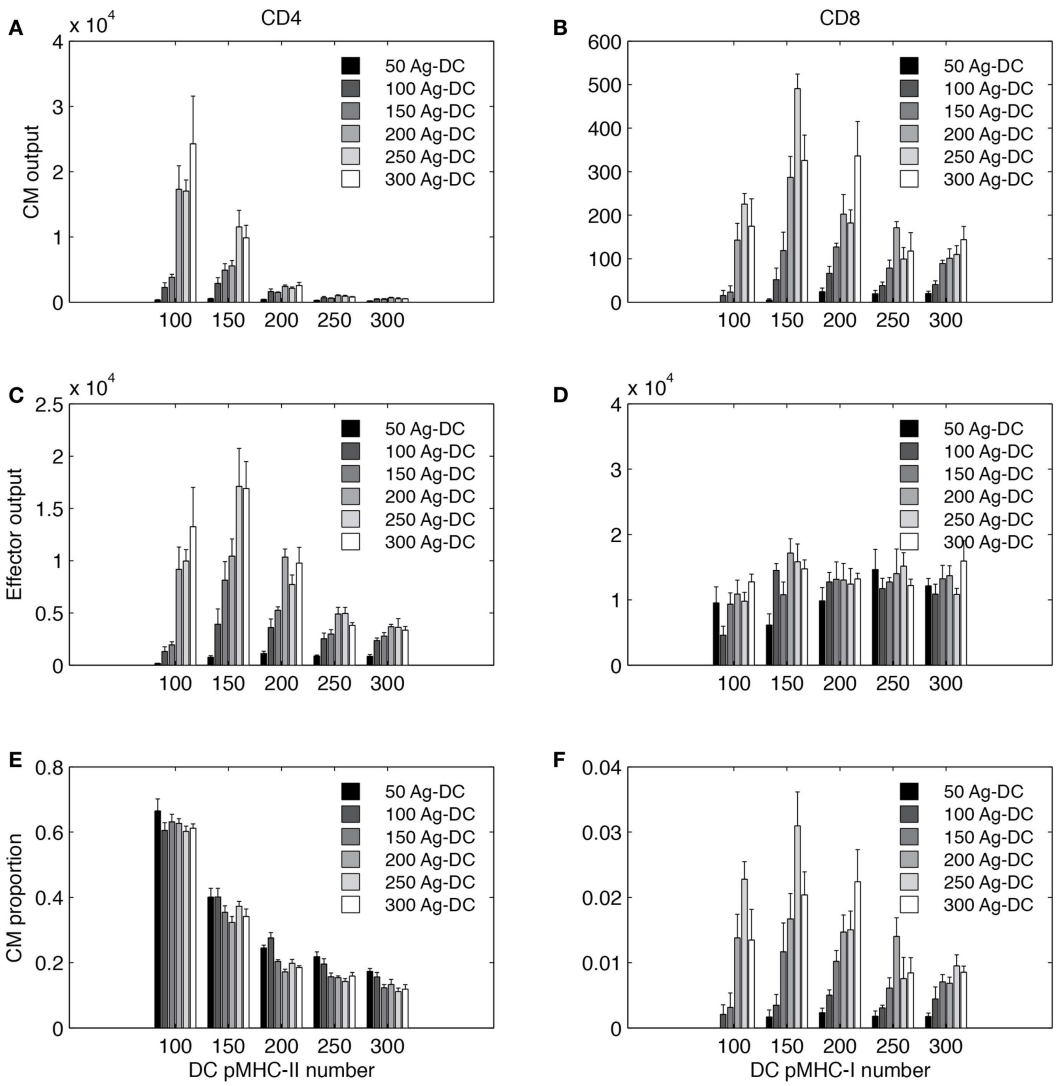


FIGURE 9 | Simulated T cell differentiation is influenced by number of DCs and number of pMHC molecules displayed. *X*-axis: number of pMHC molecules on an Ag-DC. Bar color: number of Ag-DCs recruited to the grid. **(A,B)** Number of central memory (CM) cells produced. **(C,D)** Number of effector cells produced. **(E,F)** Fraction of CM cells in the expanded population.

cells, but this benefit is diminished when pMHC molecule levels are high.

Interestingly, for each subset of cells, the effects of increased pMHC molecule levels on the surface of DCs are different. pMHC levels are always negatively correlated with CM output of CD4+ T cells (**Figure 9A**). However, the highest numbers of effector CD4+ T cell output are reached at intermediate levels of pMHC (**Figure 9C**). CD8+ T cells are affected in a different pattern. Intermediate levels of pMHC are required for higher CM production (**Figure 9B**). Our explanation for this difference is that we defined our rules based on LCMV experiments and other studies suggesting that CD8+ T cell memory establishment is dependent on DC licensing by effector CD4+ T cells, and intermediate levels of pMHC are required so that LDC numbers will not be the bottleneck for memory CD8+ T cell production. Also different than CD4+ T cells, effector CD8+ T cell output first increases with pMHC levels on DCs and then remains relatively stable (**Figure 9D**). The fraction of CM among total expanded population is shown in **Figures 9E,F**.

In general, our simulations suggest that in order to obtain high CD4+ T CM cell production, DCs with lower pMHC levels should be provided. However, DCs with high pMHC levels maximize CD8+ effector output. CD4+ effectors and CD8+ CM T cells require intermediate pMHC levels. Increased recruitment of Ag-DC boosts all four subsets to different extents.

SENSITIVITY ANALYSIS DETECTS MECHANISMS CORRELATED WITH STRENGTH OF RECALL RESPONSES

We also studied how other mechanisms, such as thresholds of different checkpoints in the LN, conversion rates of EM to CM, and

recruitment rate to sites of infection from the blood, shape model outputs (see Table S3 in Supplementary Material for parameters varied. The mechanisms they control are explained in the column "description"). We performed intra- and inter-compartment sensitivity analysis (see Materials and Methods), and PRCCs were calculated to assess the monotonic correlation between values of the parameters governing these mechanisms, and outputs including: LN production and blood concentration of effector and CM cells in both primary and recall responses (**Tables 1–3**).

We found that production of effector cells in primary responses is most sensitive to the following mechanisms: binding time (negatively correlated, **Tables 1** and **2**), the probability of effectors differentiating into EM (negatively correlated, **Tables 1** and **2**), and extra recruitment in inflammation (positively correlated, **Tables 1** and **2**). In addition, CD4+ T cell effector output negatively correlates with the stimulation threshold for priming but positively correlates with the threshold for differentiation into effectors (**Tables 1** and **2**); CD8+ T cell effector cell output negatively correlates with extra recruitment of CD4+ T cells (**Tables 1** and **2**). Generation of CM cells during a primary response is sensitive to a similar set of mechanisms, along with some additional ones: CD4+ T cell CM output is negatively affected by the efficiency of CM cells to accumulate stimulation signal; CD8+ T cell CM output is positively correlated with threshold to become effectors (**Tables 1** and **2**). Interestingly, in the recall response, the mechanisms to which effector cell production are sensitive are consistent with that of those affecting effector cells and CM cells during primary responses. For instance, the LN output of effectors in the recall response has a more significant negative correlation with EM differentiation probability than in the primary response, blood CM concentration. This is in accordance with the intuition that a strong recall response requires both memory generation during a primary response and priming efficiency during secondary challenge.

We then perform an inter-compartment sensitivity analysis by comparing how the readouts for the same cell types from both LN and blood are affected by corresponding LN and blood mechanisms. LN mechanisms affect both LN and blood outputs in similar ways, but only blood outputs are sensitive to blood mechanisms (**Table 3**). The recruitment rates of effector cells to

Table 1 | PRCC results: tracking sensitivity of outputs of LN cells to LN mechanisms.

Primary			Recall		
Mechanism	**PRCC**	**Significance**	**Mechanism**	**PRCC**	**Significance**
CD4+ EFFECTOR GENERATED FROM LN					
Bind time	−0.90	−−−	Bind time	−0.82	−−−
Probability EM	−0.73	−−−	Probability EM	−0.64	−−−
Priming checkpoint threshold	−0.42	−−−	Priming checkpoint threshold	−0.40	−−−
CM bind time	−0.29	−−	CM bind time	−0.37	−−−
DC licensing probability	−0.26	−−	DC licensing probability	−0.29	−−
Extra recruitment	0.38	+++	Efficiency CM	−0.27	−−
Effector checkpoint threshold	0.67	+++	Extra recruitment	0.35	+++
			Effector checkpoint threshold	0.61	+++
CD4+ CM GENERATED FROM LN					
Bind time	−0.88	−−−	Bind time	−0.62	−−−
Priming checkpoint threshold	−0.70	−−−	Priming checkpoint threshold	−0.48	−−−
Efficiency CM	−0.29	−−	CM bind time	−0.44	−−−
CM bind time	−0.26	−−	Efficiency CM	−0.33	−−−
Extra recruitment	0.33	+++	Extra recruitment	0.26	++
Effector checkpoint threshold	0.81	+++	Effector checkpoint threshold	0.61	+++
CD8+ EFFECTOR GENERATED FROM LN					
Extra recruitment (CD4)	−0.56	−−−	Probability EM	−0.74	−−−
Bind time	−0.41	−−−	Extra recruitment (CD4)	−0.56	−−−
DC licensing prob.	−0.25	−−	Bind time (CD8)	−0.37	−−−
Probability EM	−0.25	−−	DC licensing probability	−0.21	−−
CD8+ CM GENERATED FROM LN					
Bind time	−0.71	−−−	Efficiency CM	−0.42	−−−
Priming checkpoint threshold	−0.71	−−−	CM bind time	−0.35	−−−
CM bind time	−0.19	−	Priming checkpoint threshold	−0.22	−
Extra recruitment	0.30	++	Bind time	−0.18	−
Effector checkpoint threshold	0.77	+++	Prob. EM	0.33	+++
			Effector checkpoint threshold	0.35	+++

−/+, $p < 0.001$, with negative or positive correlation; −−/++, $p < 10^{-6}$, with negative or positive correlation; −−−/+++, $p < 10^{-9}$, with negative or positive correlation.

Table 2 | PRCC results: tracking sensitivity of concentrations of cells in Blood to LN mechanisms.

Primary			Recall		
Mechanism	**PRCC**	**Significance**	**Mechanism**	**PRCC**	**Significance**
CD4+ EFFECTOR CONCENTRATION					
Bind time	−0.83	−−−	Bind time	−0.71	−−−
Probability EM	−0.74	−−−	Probability EM	−0.67	−−−
Priming checkpoint threshold	−0.42	−−−	DC licensing probability	−0.39	−−−
DC licensing probability	−0.41	−−−	CM bind time	−0.38	−−−
CM bind time	−0.29	−−	Priming checkpoint threshold	−0.35	−−−
Extra recruitment	0.32	+++	Efficiency CM	−0.25	−−
Effector checkpoint threshold	0.64	+++	Extra recruitment	0.27	++
			Effector checkpoint threshold	0.53	+++
CD4+ CM CONCENTRATION					
Bind time	−0.89	−−−	Bind time	−0.70	−−−
Priming checkpoint threshold	−0.71	−−−	Priming checkpoint threshold	−0.53	−−−
Efficiency CM	−0.30	−−	CM bind time	−0.43	−−−
CM bind time	−0.26	−−	Efficiency CM	−0.33	−−−
Extra recruitment	0.33	+++	Extra recruitment	0.26	++
Effector checkpoint threshold	0.81	+++	Effector checkpoint threshold	0.66	+++
CD8+ EFFECTOR CONCENTRATION					
Extra recruitment (CD4+)	−0.57	−−−	Probability EM	−0.72	−−−
Bind time	−0.31	−−−	Extra recruitment (CD4+)	−0.55	−−−
DC licensing probability	−0.28	−−	Bind time (CD8+)	−0.28	−−
Probability EM	−0.27	−−	DC licensing probability	−0.23	−
			Bind time (CD4+)	0.18	+
CD8+ CM CONCENTRATION					
Probability EM	−0.78	−−−	Probability EM	−0.82	−−−
Bind time	−0.53	−−−	Extra recruitment (CD4+)	−0.42	−−−
Priming checkpoint threshold	−0.39	−−−	Bind time	−0.38	−−−
Extra recruitment (CD4+)	−0.37	−−−	Efficiency CM	−0.30	−−
Effector checkpoint threshold	0.52	+++	CM bind time	−0.27	−−
			Priming checkpoint threshold	−0.25	−−
			Effector checkpoint threshold	0.39	+++

−/+, $p < 0.001$, with negative or positive correlation; −−/++, $p < 10^{-6}$, with negative or positive correlation; −−−/+++, $p < 10^{-9}$, with negative or positive correlation.

Table 3 | PRCC results: tracking sensitivity of concentrations of cells in blood to blood mechanisms.

Primary			Recall		
Mechanism	**PRCC**	**Significance**	**Mechanism**	**PRCC**	**Significance**
CD4+ EFFECTOR CONCENTRATION					
Recruit to site of infection (ξ_{E4})	−0.79	−−−	Recruit to site of infection (ξ_{E4})	−0.54	−−
CD4+ CM					
Recruit to site of infection (ξ_{EM4})	−0.39	−	Probability EM	0.44	+
Probability EM	−0.80	+++			
CD8+ EFFECTOR CONCENTRATION					
Recruit to site of infection (ξ_{E8})	−0.39	−	Recruit to site of infection (ξ_{E8})	−0.43	−
			Recruit to site of infection (ξ_{EM8})	−0.36	−
			Probability EM	0.72	+++
CD8+ CM CONCENTRATION					
Recruit to site of infection (ξ_{EM8})	−0.41	−	Recruit to site of infection (ξ_{EM8})	−0.45	−
Probability EM	0.93	+++	Probability EM	0.83	+++

−/+, $p < 0.001$, with negative or positive correlation; −−/++, $p < 10^{-6}$, with negative or positive correlation; −−−/+++, $p < 10^{-9}$, with negative or positive correlation.

sites of infection reduce blood effector levels. Conversion rates from EM to CM induce both blood CD4+ and CD8+ T cell CM levels. Also, our predictions suggest that dynamics occurring in blood do not significantly affect dynamics occurring in LNs during both primary or recall responses, as no significant correlation is detected.

DISCUSSION

Single cognate naïve T cells have been known to be able to generate both memory and effector progenies. Moreover, recent studies further demonstrated that the fate of identical naïve cells is heterogeneous. By understanding which mechanisms contribute to this heterogeneity and in which way they are contributing, it is possible to manipulate the priming environment so that differentiation of activated precursor cells can be routed to favor generation of a desired population toward specific needs. For example, in the context of vaccination, establishing a significant antigen-specific CM population has a high priority. On the other hand, massive production of effector cells could be the key issue considered when using immunotherapies against active diseases. Our new findings suggest that pMHC number on the surface of APCs provides such a handle; and using our model we can enhance the production of specific T cell types (CD4+/CD8+, effector/memory) in different ideal ranges. By fitting our model to data collected from experiments designed for a specific antigen, we will be capable of making quantitative predictions of DC stimulation levels that maximize the generation of particular subsets of T cells, which are most relevant to the circumstances.

In order to study how mechanisms in LN and blood influence the generation of effector and memory T cells, we developed a hybrid model with both LN and blood compartments to simulate immune responses in both primary and recall challenges. Using this model, we generated T cell dynamics in blood and LN during infections that are similar to murine models (61–63) and also can capture heterogeneous differentiation observations of individual cognate naïve T cells (18, 19). Furthermore, our model predicts that the outputs of different subsets of T cells that arise during immune responses, including effector and memory, CD4+, and CD8+ T cells, respond differently to the amount of stimulation they receive from antigen-bearing DCs during priming. Simulations showed that CD4+ CM T cell generation is maximized at low pMHC-II levels, and intermediate pMHC-II levels result in the highest number of effector CD4+ T cell generation. However, further increases in pMHC-II levels reduce generation of both effector and CM CD4+ T cells. On the other hand, intermediate pMHC-I level is required to generate the highest levels of CD8+ CM cells, and high pMHC-I level favors CD8+ T effector cell generation.

Results from our previous study using a 2D model showed pMHC levels always compensate for DC numbers to induce effector T cell production (25) in a trade-off fashion. We find a similar trend herein for CD8+ T cells, and for CD4+ T cells, when DC numbers are small. But when DC numbers are large, high pMHC levels are playing an opposite (for CD4+ T cell) or insignificant (for CD8+ T cell) role. This can be explained by the findings from our 3D LN model (30), that DC searching time for T cells is far more efficient in a 3D model environment than 2D. Thus, even for high total DC numbers in the 2D study, there are likely insufficient DCs, suggesting what we observed in the 2D model represents only the case with relatively low DC numbers in 3D.

While our model is able to make some important predictions, further development to include more detail regarding events during antigen presentation is called for. First, DCs are known to be a heterogeneous population, with subsets of cells diversified in origin and function. Different DC subsets are differentially involved in T cell priming. For example, lymphoid organ-resident DCs are specialized for cross-presentation, while inflammatory DCs stimulate T_H17 polarization (70, 71). Furthermore, the stimulation that T cells receive from DCs are also combinations of multiple signals, including TCR avidity, co-stimulation regulation, and environmental, such as inflammatory cytokine profiles. Reducing these signals to a general stimulation signal package represented with pMHC levels as done here helps conceptualize the question and generate theoretical insights; nonetheless, adding these details will confer power for predicting more precise manipulations of the immune response. Harnessing the power of both mathematical and computational modeling and wet-lab investigation, our systems biology approach can eventually provide guidance for clinical practices in an era of personalized medicine.

ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. This research was supported by the following grants: NIH R01 HL106804 (awarded to Denise Kirschner), R01 EB012579 (awarded to Denise Kirschner and Jennifer J. Linderman), and R01 HL 110811 (awarded to Denise Kirschner and Jennifer J. Linderman) and by a University of Michigan Rackham International Student Fellowship awarded to Chang Gong.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00057/abstract

REFERENCES

1. Randolph GJ, Angeli V, Swartz MA. Dendritic-cell trafficking to lymph nodes through lymphatic vessels. *Nat Rev Immunol* (2005) **5**:617–28. doi:10.1038/nri1670
2. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. *Science* (1999) **286**:958–61. doi:10.1126/science.286.5441.958
3. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the αβ TCR repertoire of naive mouse splenocytes. *J Immunol* (2000) **164**:5782–7.
4. Blattman JN, Antia R, Sourdive DJ, Wang X, Kaech SM, Murali-Krishna K, et al. Estimating the precursor frequency of naive antigen-specific CD8 T cells. *J Exp Med* (2002) **195**:657–64. doi:10.1084/jem.20001021
5. Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DA, et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* (1992) **358**:764–8. doi:10.1038/358764a0

6. Butcher EC, Picker LJ. Lymphocyte homing and homeostasis. *Science* (1996) **272**:60–6. doi:10.1126/science.272.5258.60
7. Banchereau J, Steinman RM. Dendritic cells and the control of immunity. *Nature* (1998) **392**:245–52. doi:10.1038/32588
8. Girard JP, Moussion C, Forster R. HEVs, lymphatics and homeostatic immune cell trafficking in lymph nodes. *Nat Rev Immunol* (2012) **12**:762–73. doi:10.1038/nri3298
9. Masopust D, Schenkel JM. The integration of T cell migration, differentiation and function. *Nat Rev Immunol* (2013) **13**:309–20. doi:10.1038/nri3442
10. Bajenoff M, Granjeaud S, Guerder S. The strategy of T cell antigen-presenting cell encounter in antigen-draining lymph nodes revealed by imaging of initial T cell activation. *J Exp Med* (2003) **198**:715–24. doi:10.1084/jem.20030167
11. Mempel TR, Henrickson SE, Von Andrian UH. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature* (2004) **427**:154–9. doi:10.1038/nature02238
12. Grigorova IL, Panteleev M, Cyster JG. Lymph node cortical sinus organization and relationship to lymphocyte egress dynamics and antigen exposure. *Proc Natl Acad Sci U S A* (2010) **107**:20447–52. doi:10.1073/pnas.1009968107
13. Ahmed R, Gray D. Immunological memory and protective immunity: understanding their relation. *Science* (1996) **272**:54–60. doi:10.1126/science.272.5258.54
14. Sallusto F, Lenig D, Forster R, Lipp M, Lanzavecchia A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* (1999) **401**:708–12. doi:10.1038/44385
15. Cyster JG. Chemokines, sphingosine-1-phosphate, and cell migration in secondary lymphoid organs. *Annu Rev Immunol* (2005) **23**:127–59. doi:10.1146/annurev.immunol.23.021704.115628
16. Mackay CR. Homing of naive, memory and effector lymphocytes. *Curr Opin Immunol* (1993) **5**:423–7. doi:10.1016/0952-7915(93)90063-X
17. Dutton RW, Bradley LM, Swain SL. T cell memory. *Annu Rev Immunol* (1998) **16**:201–23. doi:10.1146/annurev.immunol.16.1.201
18. Buchholz VR, Flossdorf M, Hensel I, Kretschmer L, Weissbrich B, Graf P, et al. Disparate individual fates compose robust CD8+ T cell immunity. *Science* (2013) **340**:630–5. doi:10.1126/science.1235454
19. Gerlach C, Rohr JC, Perie L, Van Rooij N, Van Heijst JW, Velds A, et al. Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* (2013) **340**:635–9. doi:10.1126/science.1235487
20. Antia R, Ganusov VV, Ahmed R. The role of models in understanding CD8+ T-cell memory. *Nat Rev Immunol* (2005) **5**:101–11. doi:10.1038/nri1550
21. De Boer RJ, Perelson AS. Quantifying T lymphocyte turnover. *J Theor Biol* (2013) **327**:45–87. doi:10.1016/j.jtbi.2012.12.025
22. Beltman JB, Maree AF, De Boer RJ. Spatial modelling of brief and long interactions between T cells and dendritic cells. *Immunol Cell Biol* (2007) **85**:306–14. doi:10.1038/sj.icb.7100054
23. Riggs T, Walts A, Perry N, Bickle L, Lynch JN, Myers A, et al. A comparison of random vs. chemotaxis-driven contacts of T cells with dendritic cells during repertoire scanning. *J Theor Biol* (2008) **250**:732–51. doi:10.1016/j.jtbi.2007.10.015
24. Bogle G, Dunbar PR. T cell responses in lymph nodes. *Wiley Interdiscip Rev Syst Biol Med* (2010) **2**:107–16. doi:10.1002/wsbm.47
25. Linderman JJ, Riggs T, Pande M, Miller M, Marino S, Kirschner DE. Characterizing the dynamics of CD4+ T cell priming within a lymph node. *J Immunol* (2010) **184**:2873–85. doi:10.4049/jimmunol.0903117
26. Bogle G, Dunbar PR. On-lattice simulation of T cell motility, chemotaxis, and trafficking in the lymph node paracortex. *PLoS One* (2012) **7**:e45258. doi:10.1371/journal.pone.0045258
27. Vroomans RM, Maree AF, De Boer RJ, Beltman JB. Chemotactic migration of T cells towards dendritic cells promotes the detection of rare antigens. *PLoS Comput Biol* (2012) **8**:e1002763. doi:10.1371/journal.pcbi.1002763
28. Kirschner DE, Linderman JJ. Mathematical and computational approaches can complement experimental studies of host-pathogen interactions. *Cell Microbiol* (2009) **11**:531–9. doi:10.1111/j.1462-5822.2008.01281.x
29. Mirsky HP, Miller MJ, Linderman JJ, Kirschner DE. Systems biology approaches for understanding cellular mechanisms of immunity in lymph nodes during infection. *J Theor Biol* (2011) **287**:160–70. doi:10.1016/j.jtbi.2011.06.037
30. Gong C, Mattila JT, Miller M, Flynn JL, Linderman JJ, Kirschner D. Predicting lymph node output efficiency through systems biology. *J Theor Biol* (2013) **335**:169–84. doi:10.1016/j.jtbi.2013.06.016
31. Miller MJ, Wei SH, Parker I, Cahalan MD. Two-photon imaging of lymphocyte motility and antigen response in intact lymph node. *Science* (2002) **296**:1869–73. doi:10.1126/science.1070051
32. Constant S, Pfeiffer C, Woodard A, Pasqualini T, Bottomly K. Extent of T cell receptor ligation can determine the functional differentiation of naive CD4+ T cells. *J Exp Med* (1995) **182**:1591–6. doi:10.1084/jem.182.5.1591
33. Lanzavecchia A, Sallusto F. Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells. *Science* (2000) **290**:92–7. doi:10.1126/science.290.5489.92
34. Gett AV, Sallusto F, Lanzavecchia A, Geginat J. T cell fitness determined by signal strength. *Nat Immunol* (2003) **4**:355–60. doi:10.1038/ni908
35. Kaech SM, Cui W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* (2012) **12**:749–61. doi:10.1038/nri3307
36. Viola A, Lanzavecchia A. T cell activation determined by T cell receptor number and tunable thresholds. *Science* (1996) **273**:104–6. doi:10.1126/science.273.5271.104
37. Itoh Y, Germain RN. Single cell analysis reveals regulated hierarchical T cell antigen receptor signaling thresholds and intraclonal heterogeneity for individual cytokine responses of CD4+ T cells. *J Exp Med* (1997) **186**:757–66. doi:10.1084/jem.186.5.757
38. Iezzi G, Karjalainen K, Lanzavecchia A. The duration of antigenic stimulation determines the fate of naive and effector T cells. *Immunity* (1998) **8**:89–95. doi:10.1016/S1074-7613(00)80461-6
39. Lanzavecchia A, Sallusto F. Progressive differentiation and selection of the fittest in the immune response. *Nat Rev Immunol* (2002) **2**:982–7. doi:10.1038/nri959
40. Gett AV, Hodgkin PD. A cellular calculus for signal integration by T cells. *Nat Immunol* (2000) **1**:239–44. doi:10.1038/79782
41. Jacob J, Baltimore D. Modelling T-cell memory by genetic marking of memory T cells in vivo. *Nature* (1999) **399**:593–7. doi:10.1038/21208
42. Opferman JT, Ober BT, Ashton-Rickardt PG. Linear differentiation of cytotoxic effectors into memory T lymphocytes. *Science* (1999) **283**:1745–8. doi:10.1126/science.283.5408.1745
43. Pepper M, Jenkins MK. Origins of CD4(+) effector and central memory T cells. *Nat Immunol* (2011) **12**:467–71. doi:10.1038/ni.2038
44. Byrne JA, Butler JL, Cooper MD. Differential activation requirements for virgin and memory T cells. *J Immunol* (1988) **141**:3249–57.
45. Bachmann MF, Gallimore A, Linkert S, Cerundolo V, Lanzavecchia A, Kopf M, et al. Developmental regulation of Lck targeting to the CD8 coreceptor controls signaling in naive and memory T cells. *J Exp Med* (1999) **189**:1521–30. doi:10.1084/jem.189.10.1521
46. Wiesel M, Oxenius A. From crucial to negligible: functional CD8(+) T-cell responses and their dependence on CD4(+) T-cell help. *Eur J Immunol* (2012) **42**:1080–8. doi:10.1002/eji.201142205
47. Stemberger C, Huster KM, Koffler M, Anderl F, Schiemann M, Wagner H, et al. A single naive CD8+ T cell precursor can develop into diverse effector and memory subsets. *Immunity* (2007) **27**:985–97. doi:10.1016/j.immuni.2007.10.012
48. Kaech SM, Ahmed R. Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naïve cells. *Nat Immunol* (2001) **2**:415–22. doi:10.1038/87720
49. Hawkins ED, Markham JF, McGuinness LP, Hodgkin PD. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proc Natl Acad Sci U S A* (2009) **106**:13457–62. doi:10.1073/pnas.0905629106
50. Chang JT, Palanivel VR, Kinjyo I, Schambach F, Intlekofer AM, Banerjee A, et al. Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science* (2007) **315**:1687–91. doi:10.1126/science.1139393
51. Bajaria SH, Webb G, Cloyd M, Kirschner D. Dynamics of naive and memory CD4+ T lymphocytes in HIV-1 disease progression. *J Acquir Immune Defic Syndr* (2002) **30**:41–58. doi:10.1097/00126334-200205010-00006
52. Steinmann GG, Klaus B, Muller-Hermelink HK. The involution of the ageing human thymic epithelium is independent of puberty. A morphometric study. *Scand J Immunol* (1985) **22**:563–75. doi:10.1111/j.1365-3083.1985.tb01916.x
53. Homann D, Teyton L, Oldstone MB. Differential regulation of antiviral T-cell immunity results in stable CD8+ but declining CD4+ T-cell memory. *Nat Med* (2001) **7**:913–9. doi:10.1038/90950
54. Wherry EJ, Teichgraber V, Becker TC, Masopust D, Kaech SM, Antia R, et al. Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nat Immunol* (2003) **4**:225–34. doi:10.1038/ni889
55. Fallahi-Sichani M, El-Kebir M, Marino S, Kirschner DE, Linderman JJ. Multi-scale computational modeling reveals a critical role for TNF-alpha receptor 1

dynamics in tuberculosis granuloma formation. *J Immunol* (2011) **186**:3472–83. doi:10.4049/jimmunol.1003299
56. Marino S, El-Kebir M, Kirschner D. A hybrid multi-compartment model of granuloma formation and T cell priming in Tuberculosis. *J Theor Biol* (2011) **280**:50–62. doi:10.1016/j.jtbi.2011.03.022
57. Fallahi-Sichani M, Kirschner DE, Linderman JJ. NF-kappaB signaling dynamics play a key role in infection control in tuberculosis. *Front Physiol* (2012) **3**:170. doi:10.3389/fphys.2012.00170
58. Sprent J, Tough DF. T cell death and memory. *Science* (2001) **293**:245–8. doi:10.1126/science.1062416
59. Prlic M, Hernandez-Hoyos G, Bevan MJ. Duration of the initial TCR stimulus controls the magnitude but not functionality of the CD8+ T cell response. *J Exp Med* (2006) **203**:2135–43. doi:10.1084/jem.20060928
60. Roederer M, Dubs JG, Anderson MT, Raju PA, Herzenberg LA. CD8 naive T cell counts decrease progressively in HIV-infected adults. *J Clin Invest* (1995) **95**:2061–6. doi:10.1172/JCI117892
61. De Boer RJ, Oprea M, Antia R, Murali-Krishna K, Ahmed R, Perelson AS. Recruitment times, proliferation, and apoptosis rates during the CD8(+) T-cell response to lymphocytic choriomeningitis virus. *J Virol* (2001) **75**:10663–9. doi:10.1128/JVI.75.22.10663-10669.2001
62. Antia R, Bergstrom CT, Pilyugin SS, Kaech SM, Ahmed R. Models of CD8+ responses: 1. What is the antigen-independent proliferation program. *J Theor Biol* (2003) **221**:585–98. doi:10.1006/jtbi.2003.3208
63. De Boer RJ, Homann D, Perelson AS. Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J Immunol* (2003) **171**:3928–35.
64. Marino S, Hogue IB, Ray CJ, Kirschner DE. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J Theor Biol* (2008) **254**:178–96. doi:10.1016/j.jtbi.2008.04.011
65. McKay MD, Beckman RJ, Conover WJ. Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* (1979) **21**:239–45. doi:10.2307/1268522
66. Erkeller-Yuksel FM, Deneys V, Yuksel B, Hannet I, Hulstaert F, Hamilton C, et al. Age-related changes in human blood lymphocyte subpopulations. *J Pediatr* (1992) **120**:216–22. doi:10.1016/S0022-3476(05)80430-5
67. Douek DC, McFarland RD, Keiser PH, Gage EA, Massey JM, Haynes BF, et al. Changes in thymic function with age and during the treatment of HIV infection. *Nature* (1998) **396**:690–5. doi:10.1038/25374
68. Ravkov EV, Williams MA. The magnitude of CD4+ T cell recall responses is controlled by the duration of the secondary stimulus. *J Immunol* (2009) **183**:2382–9. doi:10.4049/jimmunol.0900319
69. Zehn D, King C, Bevan MJ, Palmer E. TCR signaling requirements for activating T cells and for generating memory. *Cell Mol Life Sci* (2012) **69**:1565–75. doi:10.1007/s00018-012-0965-x
70. Segura E, Durand M, Amigorena S. Similar antigen cross-presentation capacity and phagocytic functions in all freshly isolated human lymphoid organ-resident dendritic cells. *J Exp Med* (2013) **210**:1035–47. doi:10.1084/jem.20121103
71. Segura E, Touzot M, Bohineust A, Cappuccio A, Chiocchia G, Hosmalin A, et al. Human inflammatory dendritic cells induce Th17 cell differentiation. *Immunity* (2013) **38**:336–48. doi:10.1016/j.immuni.2012.10.018

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; accepted: 31 January 2014; published online: 19 February 2014.
*Citation: Gong C, Linderman JJ and Kirschner D (2014) Harnessing the heterogeneity of T cell differentiation fate to fine-tune generation of effector and memory T cells. Front. Immunol. **5**:57. doi: 10.3389/fimmu.2014.00057*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Gong, Linderman and Kirschner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 27 November 2013
doi: 10.3389/fimmu.2013.00413

The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis

Adrien Six [1,2,3,4,5] *†, ***Maria Encarnita Mariotti-Ferrandiz*** [1,2,3,5], ***Wahiba Chaara*** [1,2,3,4,5], ***Susana Magadan*** [6], ***Hang-Phuong Pham*** [1,2], ***Marie-Paule Lefranc*** [7], ***Thierry Mora*** [8], ***Véronique Thomas-Vaslin*** [1,2,3,5], ***Aleksandra M. Walczak*** [9] ***and Pierre Boudinot*** [6†]

[1] *UPMC University Paris 06, UMR 7211, Immunology-Immunopathology-Immunotherapy (I3), Paris, France*
[2] *CNRS, UMR 7211, Immunology-Immunopathology-Immunotherapy (I3), Paris, France*
[3] *INSERM, UMR_S 959, Immunology-Immunopathology-Immunotherapy (I3), Paris, France*
[4] *AP-HP, Hôpital Pitié-Salpêtrière, CIC-BTi Biotherapy, Paris, France*
[5] *AP-HP, Hôpital Pitié-Salpêtrière, Département Hospitalo-Universitaire (DHU), Inflammation-Immunopathology-Biotherapy (i2B), Paris, France*
[6] *Institut National de la Recherche Agronomique, Unité de Virologie et Immunologie Moléculaires, Jouy-en-Josas, France*
[7] *IMGT®, The International ImMunoGeneTics Information System®, Institut de Génétique Humaine, UPR CNRS 1142, Université Montpellier 2, Montpellier, France*
[8] *Laboratoire de Physique Statistique, UMR8550, CNRS and Ecole Normale Supérieure, Paris, France*
[9] *Laboratoire de Physique Théorique, UMR8549, CNRS and Ecole Normale Supérieure, Paris, France*

Edited by:
Miles Davenport, University of New South Wales, Australia

Reviewed by:
Koji Yasutomo, University of Tokushima, Japan
John J. Miles, Queensland Institute of Medical Research, Australia

****Correspondence:***
Adrien Six, CNRS UMR 7211, UPMC, Immunology-Immunopathology-Immunotherapy (I3), BâtimentCervi, 83 bd de l'Hôpital, Paris F-75013, France
e-mail: adrien.six@upmc.fr

†Adrien Six and Pierre Boudinot have contributed equally to this work.

T and B cell repertoires are collections of lymphocytes, each characterized by its antigen-specific receptor. We review here classical technologies and analysis strategies developed to assess immunoglobulin (IG) and T cell receptor (TR) repertoire diversity, and describe recent advances in the field. First, we describe the broad range of available methodological tools developed in the past decades, each of which answering different questions and showing complementarity for progressive identification of the level of repertoire alterations: global overview of the diversity by flow cytometry, IG repertoire descriptions at the protein level for the identification of IG reactivities, IG/TR CDR3 spectratyping strategies, and related molecular quantification or dynamics of T/B cell differentiation. Additionally, we introduce the recent technological advances in molecular biology tools allowing deeper analysis of IG/TR diversity by next-generation sequencing (NGS), offering systematic and comprehensive sequencing of IG/TR transcripts in a short amount of time. NGS provides several angles of analysis such as clonotype frequency, CDR3 diversity, CDR3 sequence analysis, V allele identification with a quantitative dimension, therefore requiring high-throughput analysis tools development. In this line, we discuss the recent efforts made for nomenclature standardization and ontology development. We then present the variety of available statistical analysis and modeling approaches developed with regards to the various levels of diversity analysis, and reveal the increasing sophistication of those modeling approaches. To conclude, we provide some examples of recent mathematical modeling strategies and perspectives that illustrate the active rise of a "next-generation" of repertoire analysis.

Keywords: diversity analysis, immune receptors, next-generation sequencing, modeling, statistics, gene nomenclature, B cell repertoire, T cell repertoire

INTRODUCTION

T and B cell repertoires are collections of lymphocytes, each characterized by its antigen-specific receptor. The resources available to generate the potential repertoires are described by the genomic T cell receptor (TR) and immunoglobulin (IG) loci. TR and IG are produced by random somatic rearrangements of V, D, and J genes during lymphocyte differentiation. The product of the V-(D)-J joining, called the complementarity determining region 3 (CDR3) and corresponding to the signature of the rearrangement, binds the antigen and is responsible for the specificity of the recognition. During their differentiation, lymphocytes are subjected to selective processes, which lead to deletion of most auto-reactive cells, selection, export, and expansion, of mature T and B cells to the periphery. Primary IG and TR repertoires are therefore shaped to generate the available peripheral or mucosal repertoires. In addition, several different functional T and B cells subsets have been identified, with differential dynamics and antigen-specific patterns. These available repertoires are dramatically modified during antigen-driven responses especially in the inflammatory context of pathogen infections, autoimmune syndromes, and cancer to shape actual repertoires. When considering the importance of efficient adaptive immune responses to get rid of infections naturally or to avoid auto-reactive damages, but also for therapeutic purposes such as vaccination or cell therapy, one realizes the relevance of understanding how lymphocyte repertoires are selected during differentiation, from ontogeny to aging, and upon antigenic challenge. However, immune repertoires of expressed antigen receptors are

built by an integrated system of genomic recombination and controlled expression, and follow complex time-space developmental patterns. Thus, an efficient repertoire analysis requires both (1) methods that sample and describe the diversity of receptors at different levels for an acceptable cost and from a little amount of material and (2) analysis strategies that reconstitute the best multidimensional picture of the immune diversity from the partial information provided by the repertoire description as reviewed in Ref. (1). In the following sections, we summarize technologies developed over the past decades to describe lymphocyte repertoires and we present the growing number of analysis tools, evolving from basic to sophisticated statistics and modeling strategies with regards to the level of complexity of the data produced.

METHODS DEVELOPED TO DESCRIBE THE IG AND TR REPERTOIRES

B and T lymphocyte repertoires can be studied from different lymphoid tissues and at various biological levels, such as cell membrane or secreted proteins, transcripts or genes, according to the techniques used. Fluorescence microscopy or flow cytometry techniques allow to track and sort particular cell phenotypes and to quantify the expressed repertoire at the single-cell level with V subgroup-specific monoclonal antibodies. Alternatively, the IG or TR diversity may be also analyzed using proteomics methods from either the serum (for IG) or dedicated cell extracts. Finally, molecular biology techniques assess the repertoire at the genomic DNA or transcriptional levels, qualitatively and/or quantitatively.

ANALYSIS OF IG AND TR REPERTOIRES AT THE PROTEIN LEVEL

Flow cytometry single-cell repertoire analysis

The frequency of lymphocytes expressing a given IG or TR can be determined using flow cytometry when specific monoclonal antibodies are available. This technique allows for the combined analysis of the antigen receptor and of other cell surface markers. Currently, using flow cytometry, up to 13 parameters can be routinely studied at once, reaching 20 parameters with the last generation flow cytometers and 70–100 parameters with mass cytometry (2). Seminal studies in mice using specific anti-TRBV antibodies have led to the characterization of the central tolerance selection processes that occur in thymus (3–5). Later on, a comprehensive description of the human TRBV repertoire was setup (6), when monoclonal antibodies became available for most of the TRBV subgroups. Repertoire analysis with flow cytometry provides a qualitative and quantitative analyses of the variable region, often done on heterogeneous cell populations, in order to decipher, for example, selection events related to aging, perturbations, and treatments (7). However, this technology is naturally limited by the availability of specific monoclonal antibodies, and does not address more detailed issues such as junction diversity. Furthermore, polymorphism of the IG or TR genes (8, 9) may constitute a serious limitation for a systematic survey using these approaches.

Proteomic repertoire analysis for serum immunoglobulins

Recent developments of proteomics tools now offer sensitivity levels applicable to IG repertoire analysis. Such a description at the protein level takes into account all post transcriptional and translational modifications.

PANAMA-blot technology. A semi-quantitative immunoblot, called the PANAMA-blot technique (10), allows for the identification of the antibody reactivities present in collection of sera (or cell culture supernatant) against a given source of antigens (10–12). Briefly, a selected source of antigens is subjected to preparative SDS-PAGE, transferred onto nitrocellulose membranes, then incubated with the serum to be tested allowing for the revelation of the bound antibodies using an appropriate secondary antibody coupled to alkaline phosphatase. Computer-assisted analysis of the densitometric profiles allows for the rescaling and the quantitative comparison of patterns of antibody reactivity from individuals in different groups. A large amount of data is generated when testing a range of sera against various sources of antigens. Statistical analyses are included in the PANAMA-Blot approach (as described further). This global analysis helped to reveal that the IgM repertoire in mice is selected by internal ligands and independent of external antigens (13).

This method can also lead to identify IG reactivity patterns specific for a type of pathology or clinical status and has been applied to both fundamental and clinical analysis. In particular, it was used to analyze human self-reactive antibody repertoires and their potential role for down-modulating autoimmune processes (14–16).

Antigen micro-array chips. More recently, antigen micro-array-based technology coupled to a complex two-way clustering bioinformatics analysis was developed to evaluate the serum repertoire antibodies from diabetes-prone individuals and revealed their predictive or diagnostic value. In brief, a range of antigens (proteins, peptides, nucleotides, phospholipids. . .) were plated onto glass plates and incubated with sera from individuals (human diabetes patients or mice in an experimental model of diabetes). The intensity of reactivity of the serum IG for each peptide was determined and scored against the control reactivity. Clustering analysis was then implemented to determine a potential antigen signature that significantly sorts out diabetes from non-diabetes individuals. In this way, it was found that the patterns of IgG antibodies expressed early in male NOD mice can mark susceptibility or resistance to diabetes induced later and that it is different than the pattern characteristic of healthy or diabetic mice after disease induction (17). Similarly, this clustering approach was applied in humans to successfully separate human subjects that are already diabetic from healthy people (18).

REPERTOIRE ANALYSIS AT THE GENOMIC DNA LEVEL

Other strategies that cover IG or TR repertoire analyses have been developed at the genomic DNA level. Firstly, CDR3 spectratyping studies (detailed in the following section) have been carried out at the DNA level mostly to address issues related to B or T cell development (19, 20). More recently, an original multiplex genomic PCR assay coupled to real-time PCR analysis was developed to provide a comprehensive description of the mouse T cell receptor alpha (TRA) repertoire during development (21). Although these approaches can be applied to all IG isotypes and TR, they have not been used as much as transcript CDR3 spectratyping due to sensitivity and heterozygosity issues.

Immunoglobulin or T cell receptor repertoires can also be assessed by following the diversity of rearrangement deletion circles. Since they are produced by the V-(D)-J recombination machinery when the joint signal is formed and diluted in daughter cells, they give a good representation of recently generated T or B cells. This technique has been particularly useful for describing the restoration of T cell diversity following highly active antiretroviral therapy in HIV-infected patients (22) and has been used to model thymic export (23, 24) as well as to demonstrate continued contribution of the thymus to repertoire diversity, even in older individuals (25). It also reveals that thymic output is genetically determined, and related to the extent of proliferation of T cells at DN4 stage in mice (26). However, their analysis does not provide much insight into the level of diversity since the signal joint does not vary for a given combination of genes. Therefore, the interest of such analyses is reached when combined with CDR3 spectratyping analyses to know whether a repertoire perturbation is rather attributable to newly produced T cells or peripheral T cell proliferation.

V-(D)-J JUNCTION ANALYSIS OF IG AND TR TRANSCRIPT REPERTOIRES

Original molecular-based strategies for analyzing repertoire diversity relied on cloning and hybridization of molecular probes specific for IGHV gene subgroups first by RNA colony blot assay (27). This led to the observation that IGHV gene usage is characteristic of mouse strain and is a process of random genetic combination by equiprobable expression of IGHV genes (28). The study of selection processes revealed that the IGHV region-dependent selection determines clonal persistence of B cells (29) and that selection with age leads to biased IGHV gene expression (30).

In situ hybridization on single-cells revealed that during mouse ontogeny and early development of B cells in bone marrow, there is a non-random position-dependent IGHV gene expression, favoring D-proximal IGHV gene subgroup usage (31). Thereafter, sequencing of PCR-amplified cDNA collections were obtained from samples of interest. Although fastidious, these early studies have been useful in defining the basis of human IG and TR repertoires in terms of overall distribution, CDR3-length distribution, and V-(D)-J use (32–35), sometimes leading to the identification of new IG or TR genes. Later, more practical techniques have been developed for large-scale analysis of lymphocyte repertoires, such as quantitative PCR, micro-array, and junction length spectratyping, as described below.

Quantitative RT-PCR for repertoire analysis

In parallel to qualitative CDR3 spectratyping techniques (see section below), quantitative PCR strategies were developed (36). Coupling the two techniques for all V domain-C region combinations provides a complete qualitative and quantitative picture of the repertoire (37–39) described by up to 2,000 measurements per IG isotype or TR for one sample. With the development of real-time quantitative PCR, this approach opened the possibility for a more precise evaluation of repertoire diversity (39–41). Complementary tools have been also developed in order to allow normalization of spectratype analysis such as studies by Liu et al. (42) and Mugnaini et al. (43).

Matsutani et al. (44) developed another method to quantify the expression of the human TRAV and TRBV repertoires based on hybridization with gene specific primers coated plates. The cDNA from PBMC extracted RNA are ligated to a universal adaptor which allows for a global amplification of all TRAV or TRBV cDNAs. The PCR products are then transferred onto microplates coated with oligonucleotides specific for each TRAV or TRBV regions, and the amount of hybridized material is quantified. This technique was used to analyze the TR repertoire diversity of transplanted patients (45) and adapted to the study of mouse TRAV and TRBV repertoires (46). VanderBorght et al. also developed a semi-quantitative PCR-ELISA-based method for the human TRAV and TRBV repertoire analysis (38). The combined usage of digoxigenin (DIG)-coupled nucleotides and DIG-coupled reverse TRAC or TRBC primers allowed for a quantitative measurement of the amount of amplified DNA by a sandwich ELISA.

Du et al. (47) later setup a megaplex PCR strategy to characterize the antigen-specific TRBV repertoire from sorted IFNγ-producing cells after *Mycobacterium* infection. The clonotypic TRBV PCR products were used for Taqman probes design to quantify the expression of the corresponding clonotypes from ATLAS-amplified SMART cDNAs.

Direct measurement of lymphocyte diversity using micro-arrays

Another technology, similar to the one just discussed, has been developed by the group of Cascalho et al. which allows for a direct measurement of the entire population of lymphocyte-receptors. This is accomplished by hybridization of lymphocyte-receptor specific cRNA of a lymphocyte population of interest to random oligonucleotides on a gene chip; the number of sites undergoing hybridization corresponds to the level of diversity. This method was validated and calibrated using control samples of random oligonucleotides of known diversity (1, 10^3, 10^6, 10^9) (48, 49) and successfully demonstrated that central and peripheral diversification of T lymphocytes is dependent on the diversity of the circulating IG repertoire (49, 50). Similarly, a highly sensitive micro-array-based method has been proposed to monitor TR repertoire at the single-cell level (51).

CDR3 spectratyping techniques

Immunoscope technology. Among various techniques used to analyze the T or B cell repertoires, Immunoscope, also known as CDR3 spectratyping (52, 53) consists in the analysis of the CDR3-length usage so that antigen-specific receptor repertoires can be described by thousands of measurements. In the case of naive murine repertoires, T cell populations are polyclonal and analysis typically yields eight-peak regular bell-shaped CDR3 displays (wrongly assumed to be Gaussian), each peak corresponding to a given CDR3-length. When an immune response occurs, this regular polyclonal display can be perturbed: one can see one or several prominent peaks that correspond to the oligoclonal or clonal expansion of lymphocytes. A complete description of this technique and its applications to clinical studies has been published elsewhere (54).

In the original Immunoscope publication, Cochet et al. (55) analyzed the T cell repertoire after the immunization of mice with the pigeon cytochrome c. They provided the first description of

an *ex vivo* follow-up of a primary T cell specific response in a mouse model. Their second paper analyzed the average CDR3-lengths as a function of TRBV-TRBJ combinations. In particular, the authors found a correlation between TRBV CDR1 and major histocompatibility (MH) haplotype (52). This group later published a large amount of original studies in various models such as lymphocyte development (40, 56–63), kinetics of antigen-specific responses (64–67), viral infection (68, 69), autoimmunity (70, 71), tumor-associated disease (72), and analysis of allogeneic T cell response and tolerance after transplantation (73). Notably, the combination of CDR3 spectratyping with flow cytometry-based IG or TR V frequency analysis provides a more comprehensive assessment, such as in Pilch et al. (74). For example, such an approach revealed the constriction of repertoire diversity through age-related clonal CD8 expansion (75). Similarly, a combination of CDR3 spectratyping, flow cytometry, and TR deletion circle analysis has allowed to define age-dependent incidence on thymic renewal in patients (76) or to evaluate the effects of caloric restriction in monkeys to preserve repertoire diversity (77). CDR3-length spectratyping was also used in other models, such as rainbow trout, to analyze TRB repertoire and its modifications induced by viral infection (78–80). While no tool such as monoclonal antibodies to T cell marker(s) was available in this model, this approach demonstrated that fish could mount specific T cell responses against virus, which could be found in all individuals (public clonotypes) or not (private clonotypes). Similar strategies, developed by other groups (81) and following the same approach in parallel, analyzed the IG repertoire in *Xenopus* at different stages of development, describing a more restricted IG junction diversity in the tadpole compared to the adult.

Gorski et al. (82) developed their own CDR3 spectratyping technique to analyze the complexity and stability of circulating αβT cell repertoires in patients following bone marrow transplantation as compared to normal adults. They showed that repertoire complexity of bone marrow recipients correlates with their state of immune function; in particular, individuals suffering from recurrent infections associated with T cell impairment exhibited contractions and gaps in repertoire diversity. The detailed procedure for this technique has been published in Maslanka et al. (83). A variation of this technique has been reported later by Lue et al. (84), relying on a compact glass cassette, a simpler device than the usual automated plate DNA sequencers.

Alternative technologies. Alternative CDR3 spectratyping techniques have been described such as single-strand conformation polymorphism (85–87) and heteroduplex analysis (88–91). These methods differ from the CDR3 spectratyping/Immunoscope technique mostly in the way PCR products are analyzed by performing non-denaturing polyacrylamide electrophoresis. The main advantage of these techniques is a more direct assessment of clonal expansion since PCR products migrate according to their conformation properties; therefore, presence of a predominant peak is strongly indicative of clonality when a smear migration pattern indicates polyclonality. However, these techniques have been less widely used probably because of the difficulty to make clear correlations between the expanded peaks across samples.

Another original alternative technique has been described by Bouffard et al. (92), analyzing products obtained after *in vitro* translation of PCR-amplified TR-specific products by isoelectric focusing. With this technique, clonality can also directly be assessed by looking at the obtained migration profile.

IG/TR REARRANGEMENT SEQUENCING: FROM CLONING-BASED- TO NEXT-GENERATION-SEQUENCING

In order to get a better description of IG/TR diversity at the nucleotide sequence level, thus providing fine-tuned description of the actual diversity, Sanger sequencing approaches relying on bacterial cloning of rearrangements were performed in physiological conditions globally (60, 93–99) or partially to characterize particular expansions identified by other technologies such as CDR3 spectratyping (40, 59, 100–102), flow cytometry (103). They were also used in pathological/infectious conditions (104–107) sometimes leading to antigen-specific T cell TR identification and quantification through the combination of antigen-specific T cell stimulation and cytometry-based cell sorting, anchor-PCR, and bacterial cloning-based sequencing (108).

These studies pioneered the description of the repertoire and provided fruitful information regarding the extent and modification of the diversity. However, besides being time and cost-extensive, such approaches have allowed for the analysis of 10^2–10^3 sequences, far under the estimated diversity reaching 10^6–10^7 unique clonotypes in mice and humans (40, 59, 109).

In the last decade, DNA sequencing technologies have made tremendous progresses (110) with the development of so called next-generation sequencers, already reaching four generations (111). Those instruments are designed to sequence mixtures of up to millions of DNA molecules simultaneously, instead of individual clones separately. Second generation sequencers became affordable in the last 5 years and have been used for immune repertoire analysis, starting with the seminal work of Weinstein et al. (112) where the IG repertoire of Zebrafish has been described by large-scale sequencing. Consequently, exploratory works by other groups provided an overview of the complex sequence landscape of immune repertoires in humans (113–118). More recent work aimed at addressing fundamental questions such as lineage cells commitment (119–122), generation of the diversity processes (123–125), and diversity sharing between individuals (126, 127). Finally, the power of this technology has been validated in the clinic as well (128, 129).

As seen above for other technologies, combinations of approaches have been applied to NGS. Notably, deep sequencing has been used in combination with CDR3-length spectratyping by some groups to study human (130) or rainbow trout IG (131) repertoire modifications after vaccination against bacteria or viruses. In the latter, pyrosequencing performed for relevant VH/Cμ or VH/Cτ junctions identified the clonal structure of responses, and showed, for example, that public responses are made of different clones identified by (1) distinct V-(D)-J junctions encoding the same protein sequence or (2) distinct V-(D)-J sequences differing by one or two conservative amino acid changes (131) as described for public response in mammals (132, 133). These studies showed that NGS and traditional spectratyping techniques lead to remarkably similar CDR3 distributions.

Several NGS have been developed in the past years using different sequencing technologies characterized with different speed, deepness and read length. Metzker thoroughly reviewed their principles and properties (134). Among them, three platforms, all offering benchtop sequencers with reduced cost and setup, fit with immune repertoire analysis in terms of read length and deepness. The 454/Roche platform uses pyrosequencing technology (135), which combines single nucleotide addition (SNA) with chemoluminescent detection on templates that are clonally amplified by emulsion PCR and loaded on a picotiter plate. Pyrosequencing currently has a 500 bp (GS Junior) to 700 bp (GS FLX) sequencing capacity with a respective deepness of 150,000–3,000,000 reads per run (134). The Illumina/Solexa platform technology is based on cyclic reversible termination (CRT) sequencing (an adaptation of Sanger sequencing) performed on templates clonally amplified on solid-phase bridge PCR. Protected fluorescent nucleotides are added, imaged, delabeled, and deprotected cyclically (134), providing a deeper sequencing (from 15 to 6 billion reads per run for the MiSeq to the HiSeq2500/2000) of shorter reads (100–250 bp for the very recent MiSeq) with the possibility to perform pair-end sequencing (two-side sequencing) to increase the read length after aligning the generated complementary sequences. A more recent platform, Ion Torrent/Life Technologies using an imaging free detection system may open a new era in terms of deepness (one billion reads per run) of 200 bp reads (136) in a very short time and on a benchtop sequencer. Importantly, depending on the technology, errors due to the PCR-based sample preparation and the sequencing are of major concern. Bolotin et al. (137) evaluated this issue on TR repertoire analysis of the same donor performed on the three platforms described previously; algorithms for error correction have been developed. Indeed, PCR- and sequencing-related errors represent the major concern for immune repertoire diversity analysis as they may generate artificial diversity. Illumina and 454 appear to be the most robust technologies, with Illumina having the highest throughput and 454 generating the longest reads. The currently available Ion Torrent platform, although very promising, has been shown to display the highest rate of errors in TR (137) and bacterial DNA (138) sequencing. However, such error corrections must be used with caution since they may inadvertently underestimate repertoire diversity by removing rare sequences.

With the power of such approach for genomics and transcriptomics studies in general, constant improvements are achieved to increase the sequencing deepness and read length as well as to reduce the cost, therefore offering multitude of biological explorations (139). NGS now permits a comprehensive and quantitative view of IG and TR diversity by combining and improving the sensitivity of classical approaches with accurate and large-scale sequencing. NGS has the power to identify IG or TR specific for given antigens (in combination with antigen-specific assays) and to define more complex signatures (i.e., TR sets) related to disease and/or treatment from heterogeneous T and B cell populations. Still, most of the deep sequencing efforts have been limited to only one chain of the receptor at the repertoire level (usually the β chain for TR and the heavy chain for IG). Indeed, current high-throughput approaches do not allow one to assign which combination of chains (TRA and TRB, or IGH and IGK or IGL) belong to which cell (140). A recent development by DeKosky et al. proposed a reasonably high-throughput technology to assess massively paired IG VH and VL from bulk population (141). In parallel, Turchaninova et al. (142) have proposed a similar approach for the paired analysis of the TRA and TRB chains. The parallel development of high-throughput microfluidic-based single-cell sorting will certainly push forward new developments in the field (143).

However, despite the technological advance, studies so far have mainly reported CDR3 counting and identification of major expansions. The complexity of immune repertoires is still a matter that such approach cannot completely overcome, due to the paucity of powerful analytical methods. Besides data management tools, studies are now starting to extract most of the benefit from such approach to model the immune repertoire diversity and dynamics (144), an approach that may help in understanding the interplay between cells and repertoire shaping. Accurate and powerful statistical analyses are required to manage such amount of information. Current state will be reviewed in the following sections.

POTENTIAL AND GENOMIC REPERTOIRES: A QUESTION OF ONTOLOGY AND ORTHOLOGY

Immune repertoires *sensu stricto* are expressed by lymphocyte clones, each carrying a single receptor for the antigen. Such receptors comprise IG and TR in jawed vertebrates (8, 9) and VLR in Agnathans (145). The sequences of these receptors are available in databases such as GenBank or EMBL, which are difficult to use for transversal studies due to inconsistent annotation. The IMGT® information system (see below) has largely solved this problem setting standardized gene nomenclatures, ontologies and a universal numbering of the IG/TR V and C domains, thus giving a common access to standardized data from genome, proteome, genetics, two-dimensional, and three-dimensional structures (146). The accuracy and the consistency of the IMGT® data are based on IMGT-ONTOLOGY, the first, and so far, unique ontology for immunogenetics and immunoinformatics (147).

With the development of high-throughput sequencing, large numbers of new sequences of antigen receptor genes have become available, which can be classified into different categories: genomic sequences of IG or TR (in germline configuration in genome assemblies) or fragments of IG/TR transcripts, containing the CDR3 or not. Also, these datasets can be produced from species newly sequenced, as well as from new haplotypes of well-described species.

The annotation of such sequences remains an open question. Manual annotation is not applicable, and no good automated approach has been validated yet. A relevant annotation of these massive datasets will require the integration of genomic and expression data with existing standardized description charts, as offered by IMGT®. A standardized annotation is an important issue since it facilitates the re-utilization of datasets and comparison of analyses. Thus, the description of IG and TR polymorphisms, the integration of repertoire studies with structural features of antigen-specific domains, and even the usage of new genes in genetic engineering rely on a common standard for nomenclature, numbering, and annotation (147).

To take advantage of the current standards that have been established from classical sequencing data during the last 25 years, new, fast, reliable, and human-supervised annotation methods will have to be developed, integrating directly high-throughput sequence information from the increasing number of deep sequencing platforms and technologies, at different genetic levels (genome, transcriptome, clonotype repertoires). Along this line, IMGT/HighV-QUEST offers online tools to the scientific community for the analysis of long IG and TR sequences from NGS (148).

Special attention can be paid to the orthology/paralogy relationships between similar antigen receptor genes from different species. These characteristics are essential to understand the dynamics of IG and TR loci. In fact, with many important lymphocyte subsets characterized by canonical/invariant antigen receptors, such relationships are critical to transfer functional knowledge between models. Importantly, the phylogenetic analyses required to reconstitute the evolution of antigen receptor genes are based on multiple alignments, the quality of which is highly dependent on common numbering and precise annotation of sequences.

As far as immune repertoires are characterized by the diversity of receptors specifically binding antigen/pathogen motifs to initiate a defense response, they might not be limited to lymphocyte diversifying receptors, e.g., IG, TR, and VLR. The particularity of these systems is a somatic diversification combined to a clonal structure of the repertoire, each lymphocyte clone expressing the product of a recombination/hypermutation and/or conversion process. However, many other arrays of diverse receptors binding or sensing pathogens have been discovered in metazoans, in invertebrates as well as in vertebrates.

In some cases, their diversity is really "innate," i.e., encoded in the genome as multiple genes produced by duplications. Fish NLR, finTRIMs, and NITR, primate KIR, chicken CHIR, or TLR in sea urchin, constitute good examples of such situations. While these repertoires may appear as relatively limited, polymorphism within populations, and differential expression of receptors per cell upon stimulation represent complex issues, which fall well into "traditional" repertoire approaches.

In other cases, receptors are subject to diversification processes much faster than gene duplication, which does not comply with a clonal selection pattern. The best examples are probably the DSCAM in arthropods, which hugely diversify by alternative splicing of exons encoding half-IgSF domains (149, 150), and the FREP lectins in mollusks, of which sequences are highly variable at the population level, and even between parents and offspring produced by auto-fecundation (151).

The number of such "innate" repertoires which are not expressed by clonally selected lymphocytes will likely increase with deep sequencing of new genomes/transcriptomes, as illustrated by a recent report from mussel (152). A good example of the importance of a proper structural description of key domains of receptors is provided by the extensive analysis of LRR motifs in studies on TLR evolution (153, 154). Further insights into the functions of such diverse proteins will be provided by the characterization of their expressed (available) repertoire, at different levels such as single-cells, cell populations, and animal populations.

Such analyses will require precise identification of genes and sequences as well as mutations, and a standardized approach of nomenclature and structural description will be as useful as it is for the vertebrate IG and TR sequences. Importantly, these receptors are made of a small number of structural units, such as IgSF domain or LRR domains, which suggests that standardized system(s) for sequence annotation could be developed following IMGT standards (155).

STATISTICAL ANALYSIS AND MODELING OF IMMUNE REPERTOIRE DATA

STATISTICAL REPERTOIRE ANALYSIS

The description of the repertoire modifications using flow cytometry or Immunoscope provided clear-cut and detailed insight into the clonal expansion processes during the responses against a defined antigen (64, 66). However, it is difficult to identify the relevant alterations of the repertoires in more complex situations such as pathogen infections or variable genetic backgrounds. For example, it appeared impossible to identify all significant modifications of TRB Immunoscope profiles during cerebral malaria by direct ocular comparison (107). Different methods were therefore developed to extract from IG and TR repertoire descriptions the relevant information, to encode it as numerical tables and to analyze them with statistical models.

CDR3 spectratype perturbation indices

Since the initial description of the CDR3 spectratyping technique, different scoring indices were developed or derived from the literature: "relative index of stimulation" (RIS) (55), "overall complexity score" (156), Reperturb (157), "complexity scoring system" (158), COPOM (159), Oligoscore (160), TcLandscape (161), "spectratype diversity scoring system" (162), Morisita-Horn index and Jaccard index (95–97), "absolute perturbance value" (163). A comparative review of such scoring strategies was published by Miqueu et al. (164).

In particular, the perturbation index Reperturb was developed by Gorochov et al. to perform TR repertoire analysis in HIV patient during progression to AIDS and under antiretroviral therapy. They could show drastic restrictions in the $CD8^+$ T cell repertoire at all stages of natural progression that persisted during the first 6 months of treatment. In contrast, $CD4^+$ T cell repertoire perturbations correlated with progression to AIDS with a return to a diversified repertoire in good responders to treatment (157).

Soulillou et al. refined this approach by combining the qualitative information obtained with usual CDR3 spectratyping with quantitative information of TRBV usage obtained by real-time quantitative PCR. They devised a four-dimension representation that represents TRBV subgroups, CDR3-length and percentage of TRBV use on three axis chart in addition to a color-coded representation of the CDR3 profile perturbation. Using this original approach, they were able to show that graft rejection is associated with a vigorous polyclonal accumulation of TRBV mRNA among graft-infiltrating T lymphocytes, whereas in tolerated grafts T cell repertoire is strongly altered (161, 165). Their study puts the emphasis on the importance of not only qualitative but also quantitative analysis of lymphocyte repertoires.

Platforms for repertoire data management and statistical analysis

Several platforms have been developed and rely mostly on CDR3 spectratyping and sequencing data, with recent developments to manage and analyze NGS data.

The ISEApeaks strategy and software were developed in order to satisfy the needs for efficient automated electrophoresis data retrieval and management (160, 166). ISEApeaks extracts peak area and length data generated by software used to determine fragment intensity and size. CDR3 spectratype raw data, consisting of peak areas and nucleotide lengths for each V-(D)-J-C combination, is extracted, smoothed, managed, and analyzed. The repertoires of different samples are gathered in a peak database and CDR3 spectratypes can be analyzed by different perturbation indices and multivariate statistical methods implemented in ISEApeaks. We have applied our ISEApeaks strategy in several studies. In an experimental model of cerebral malaria, we established a correlation between the quality of TR repertoire alterations and the clinical status of infected mice, whether they developed cerebral malaria or not (107). We contributed to the characterization of the membrane-associated *Leishmania* antigens (MLA) that stimulates a large fraction of naive CD4 lymphocytes. Repertoire analyses showed that MLA-induced T cell expansions used TR with various TRBV rearrangements and CDR3 lengths, a feature closer to that of polyclonal activators than of a classic antigen (167). We also revealed repertoire age-related perturbations in mice (7). ISEApeaks functions for statistical analysis was successfully applied to analyze the TR repertoire in fish as shown by our detailed analysis of the TRB repertoire of rainbow trout IELs, performed in both naive and virus-infected animals. Rainbow trout IEL TRBV transcripts were highly diverse and polyclonal in adult naive individuals, in sharp contrast with the restricted diversity of IEL oligoclonal repertoires described in birds and mammals (102). More recently, our study of the $CD8^+$ and $CD8^-$ αβ T cell repertoire suggests different regulatory patterns of those T cell patterns in fish and in mammals (168). ISEApeaks was also used to implement a new statistically based strategy for quantification of repertoire diversity (159).

Kepler et al. described another original statistical approach for CDR3 spectratype analysis, using complex procedures for testing hypotheses regarding differences in antigen receptor distribution and variable repertoire diversity in different treatment groups. This approach is based on the derivation of probability distributions directly from spectratype data instead of using *ad hoc* measures of spectratype differences (169). A software (called SpA) implementing this method has been developed and made available online (170). This approach has been used in a longitudinal analysis of TRBV repertoire during acute GvHD after stem cell transplantation (171).

Another group (163) reported the development of a new software platform, REPERTOIRE, which allows handling of CDR3 spectratyping data. This software implements a perturbation index based upon an expected normal Gaussian distribution of CDR3 length profiles.

Owing to the complexity and diversity of the immune system, immunogenetics represents one of the greatest challenges for data interpretation: a large biological expertise, a considerable effort of standardization, and the elaboration of an efficient system for the management of the related knowledge were required. To answer that challenge, IMGT®, the international ImMunoGeneTics information system®(http://www.imgt.org), was created in 1989 by one of the authors (146). Overtime, it developed standards that, since 1995, have been endorsed by the World Health Organization-International Union of Immunological Societies (WHO-IUIS) Nomenclature Committee and by the WHO-International Nonproprietary Names (INN) (172–175). IMGT® comprises seven databases (sequence, gene, and structure databases), 17 online tools and more than 15,000 pages of web resources. Among the databases, IMGT/LIGM-DB, the database for nucleotide sequences (170,685 sequences from 335 species as of July 2013) and IMGT/GENE-DB, the gene database (3,081 genes and 4,687 alleles) are of great interest for repertoire analysis. Freely available since 1997, IMGT/V-QUEST is an integrated system for the standardized analysis of collections of IG and TR rearranged nucleotide sequences (176, 177). A high-throughput version, IMGT/HighV-QUEST (148), has been released in 2010 for the analysis of long IG and TR sequences from NGS using the 454 Life Sciences technology. In the same line, other analysis tools are becoming available showing the renewed interest for repertoire analyses and modeling consecutive to NGS technology developments (178–181).

Altogether, these efforts highlight the relevance of developing more efficient and powerful technologies for the evaluation of repertoire diversity. Notably, two successful French biotech companies (TcLand, Nantes; ImmunID, Grenoble) were created in the field of repertoire analysis, using different technologies. In collaboration with ImmunID, we have proposed a novel strategy for statistical modeling of T lymphocyte repertoire data obtained in humans and humanized mice. With this model, we revealed that half of the human TRB repertoire, in terms of proportion of TRBV-TRBJ combinations, is genetically determined, the other half occurring stochastically (182). In addition, the biotechnology company "Adaptive" and the "Repertoire 10K (R10K) Project" have been recently founded by researchers respectively from the Fred Hutchinson Cancer Research Center (Seattle and Washington) and the HudsonAlpha Institute (Huntsville). Both have developed platforms (immunoSEQ®, iRepertoire®) providing researchers with a global analysis of the T or B cell receptor sequence repertoires (183). However, despite the power of this technology, studies are still limited by the ability to process the complexity of the information provided. Specific software developments for the automatic treatment and annotation of IG and TR sequences and the statistical modeling of repertoire diversity can still be improved.

Multivariate analysis

As mentioned above, the PANAMA-Blot technique also includes statistical analysis of the data. Multi-parametric analysis was introduced to compare the global reactivity of antibodies of different individuals in different groups with a given antigenic extract. This analysis has been successfully implemented to identify reactivity patterns specific for a given pathology or clinical status (10–12, 14, 15, 184). Similarly, multi-parametric analysis was also applied to TRBV spectratype analysis in an experimental cerebral malaria model (107).

Hierarchical clustering or classification algorithms have become very popular with the growing of micro-array-based transcriptome analysis. Although still uncommon for immune repertoire analysis, such approaches have been employed to categorize large sets of repertoire data without *a priori* (17, 102, 107).

Diversity indices

The concept of immune repertoire has been devised to describe the diversity of cells involved in the immune system of an individual (1). As described above, different scoring systems were developed to assess this diversity, some are heuristics but others have been borrowed from theoretical ecology and evolution. As reviewed by Magurran (185), the Shannon entropy, introduced by Claude Shannon in 1948 for the information theory, is the most used because it not only integrates the number of different species but also the relative proportion of each of these species. In 1961, Alfred Rényi generalized this entropy to a family of functions, like Species Richness, Simpson, Quadratic, and Berger–Parker indices, for quantifying the diversity, the uncertainty or randomness of a system. Most of these indices are implemented in the free software application Estimates (http://purl.oclc.org/estimates) (186). Altogether, these diversity indices constitute a collection of tools with their own sensitivity to the variety and the relative abundances of the species that are perfectly suitable for assessing immune repertoire diversity. Indeed, the very famous index of variability proposed by Kabat and Wu (187) corresponds to the ratio of Species Richness and Berger–Parker indices. In 1990, Jores et al. showed that the resolving power of this Wu-Kabat variability coefficient can be enhanced by increasing the weight on the frequency distribution of the amino acids in the formula (188). This approach inspired Stewart et al. (189) to use the Shannon entropy to demonstrate that TR amino acid composition is significantly more diverse than that of IG. In the same way, CDR3 spectratyping data can be analyzed using the relative abundance of each peak within CDR3 length global distribution. By doing so, we adjusted the original Shannon entropy, making it reaching its maximum for a Gaussian distribution, to compare the CDR3 length diversity of splenic IgM, IgD, and IgT in infected Teleost Fish (131). Recently, the Gini index, used in ecology or economics to measure the equality of distributions, was applied to individual TR clones and compared naive and memory repertoires (190). The development of deep sequencing techniques ignited a renewed interest in IG/TR repertoire. Indeed, several studies used high-throughput analysis to describe TR repertoire of key T cell subsets in human peripheral blood (115, 126, 191). This approach assessing the repertoire diversity from the relative abundance of each species in the global distribution can be decomposed hierarchically into components attributable, respectively, to variations in TRBV-TRBJ combinations and in CDR3-length (113, 117). However, most of these studies have been limited to the counting of the observed unique clonotypes. Beside the species richness, ecology-derived indices have also been applied to assess and compare immune repertoire diversity. Föhse et al. (119) used the Morisita-Horn similarity index to compare regulatory T cell repertoires between several lymphoid organs. In addition, Simpson diversity index, associated with Shannon entropy, was used to monitor TR repertoire diversity of HIV-specific CD8 T cells during antiretroviral therapy (192) but also to quantify TR repertoire recovery in the blood after allogeneic hematopoietic stem cell transplantation (128). In the same manner, Koning et al. (193) used Shannon's and Simpson's indices to show the role for the peptide component of the peptide-MH1 complex on the molecular frontline of $CD8^+$ T cell–mediated immune surveillance, by comparing the repertoire diversity of $CD8^+$ T cell populations directed against a variety of epitopes. In parallel, using Simpson's index as a metric allowed Johnson et al. (194) to model mathematically the naïve CD4 T cell repertoire contraction with age leading them to conclude that diversity plummet observed around the age of 70 could be correlated to cell-intrinsic mutations affecting cell division rate or death.

MODELING STRATEGIES

Modeling approaches have a strong tradition in immunology, usually at the boundary with other disciplines such as physics (195). Before deep sequencing data was available, general design principles were proposed as desirable features of immune repertoires, with implications for the observed repertoire diversity and dynamics (196–198). Many efforts have involved the modeling of immune cell dynamics and the effects of antigens on repertoire diversity, using differential equations descriptions of the population dynamics (199–201). Recognition in the immune system is often studied both theoretically and experimentally by probing the dynamics of cells with a specific type of receptor with respect to infections (202). Alternatively one can look at the response of a small set of chosen receptors to a specific pathogenic challenge, or careful biochemical investigation of particular receptor/antigen pairs (203, 204). Much work has been devoted to systems-biology approaches to signal processing in immune cells, as reviewed in Germain et al. (205) and Emonet and Altan-Bonnet (206). Here we focus on approaches inspired by recent advances in sequencing technologies (112, 113, 115, 116, 125, 191, 207, 208) that have opened the way for data-driven modeling of the immune repertoires and interactions between receptors and antigen.

A common modeling approach for describing receptors at the amino acid level is to choose a relevant interaction parameter (e.g., chemical affinity or hydrophobicity) and assign it a simplified digit-string representation (209). These methods are extensions of the string model, which describes both receptor and epitopes as strings of length L, with values chosen from natural numbers, and quantify their interaction by the match between the two strings (197, 210, 211). Such quantitative, physically inspired descriptions of immune receptors, despite the arbitrary choice of interaction coordinates, have proven a valuable first step in statistically describing recognition in T cells (195, 212–215). Recently, lower hydrophilicity of regulatory vs. conventional T cells was suggested from CDR3 sequencing (216).

High-throughput sequencing of immune receptors raises specific challenges compared to traditional genomic sequencing. It is harder to distinguish sequencing errors from new polymorphisms, since no corresponding pre-existing sequence exists. One of the most interesting regions when studying diversity is the CDR3 with its many insertions and deletions added to the germline sequence. These regions are often hard to align to the genomic templates, or

with each other (217). Therefore, extra care is needed when generating and analyzing sequence data. Not all sequencing technologies are equally good for all purposes (218): while 454 sequencing gives longer reads than Illumina it is known to have a greater probability of frameshift errors. In addition, primer-dependent PCR amplification biases require that raw sequence counts be normalized using control experiments (112) in order to accurately report clone sizes, as demonstrated by spike-in experiments (219). In TR repertoire studies, this is circumvented by using 5′RACE which provides an unbiased amplification of fully rearranged sequences, as recently demonstrated for TRB V-(D)-J transcripts (191).

Despite sequencing issues, statistical algorithms are often able to extract information from the data. Many studies of diversity focus on the V, D, and J gene usage of each rearranged sequence. Algorithms and tools have been developed to rapidly identify the V, D, and J genes for massive numbers of sequences (148, 178, 181). In many cases however, the assignment of a D gene to each sequence read is unreliable if the D region is too short owing to extensive trimming. Mora et al. (217) learned from data and analyzed statistical models of the D gene flanked by its junctions. These models are based on the principle of maximum entropy and make minimal assumptions about the mechanisms of diversity – they only rely on the observed frequencies of amino acid pairs along the sequence. These models were used to describe global features of the sequence ensemble, such as the probability distribution following Zipf's law (220) – the observation that the probability of sequences is inversely proportional to their frequency-rank, or the observation of peaks of frequency in sequence landscape as possible signatures of past pathogenic challenges. Recently, the estimation of repertoire diversity and clonal size distribution were analyzed by Poisson abundance models (221) and simple bivariate-Poisson-lognormal (BPLN) parametric model for fitting and analyzing TR repertoire data was proposed (222). Similarly, network analysis of IG repertoire from Weinstein et al. study revealed the possibility to identify subgroups of individuals on the basis of IG network similarity (223).

The task of characterizing the CDR3 at the nucleotide level is made difficult by the fact that a deterministic assignment of the V-(D)-J recombination process is impossible, because any given sequence can be generated by many possible recombination processes. A previous study proposed a probabilistic model of nucleotide trimming of rearranged TR genes derived from a benchmark data set of TRA and TRG V-(D)-J junctions obtained by comparison to the germline genes in the IMGT® tools (224). Recently a statistical method based on the expectation-maximization algorithm was proposed to circumvent this issue and to extract the statistical properties of junctional diversity accurately from data (124). Applying it to human non-productive DNA sequences gave insight into a universal generation mechanism, reproducible from individual to individual. It was shown that each sequence could potentially be generated by the equivalent of ∼30 equally likely ways by convergent recombination. This method showed that the potential diversity of the recombination machinery was equivalent to $\sim 10^{14}$ *equally likely* sequences (and a practically infinite total number of possible sequences), much more than the estimated 10^{12} T cells that a single human body can hold. The frequencies of the V, D, and J genes is non-uniform, even at the level of recombination, suggesting underlying physical mechanisms at work. Ndifon et al. (125) proposed a polymer model that accounts for the likelihood of connecting given genomic fragments, giving insight into the mechanistic process.

One of the ultimate goals of deep repertoire sequencing is to find signatures of the repertoire's response to its antigenic environment. A combination of clustering methods and tree reconstruction techniques have been developed (225, 226) to identify lineages in B cells and study the response to pathogenic challenges. Statistical methods have been devised to detect and quantify the extent of antigen-driven selection acting on B cells, by analyzing the patterns of hypermutations in a Bayesian framework, with applications to deep sequencing data (227, 228).

A lot remains to be done in terms of both data-driven and small-scale models of repertoire-antigen interactions. Ultimately, a close collaboration and development of experimental techniques and models can shed light on how selection at different stages shapes the repertoire, how affinity maturation changes the diversity and the link between sequence diversity and function.

FUTURE PROSPECTS OF BIOMATHEMATICAL ANALYSIS OF REPERTOIRE DATA

One of the current challenging issues in antigen-specific repertoire analysis is the development of relevant statistical analysis strategies. Biologists are usually keen on parametric tests, such as ANOVA, *t*-test, Fischer's test, among others. However, such statistical methods assume that the inherent probability distribution of the observed variable follows a normal distribution. Rock et al. (229) described that the distribution of the TR diversity is far from following this distribution, thus they proposed the use of non-parametric tests. Nevertheless, different groups are dealing with this issue in order to determine the relevant way to analyze repertoire diversity data and to propose new biostatistics strategies, including principal component analysis, discriminant analysis, hierarchical clustering, specific statistics (164, 169).

In fact, the traditional use of statistics in biology aims at the falsification of a defined hypothesis, i.e., at validating significant differences between defined situations. The recent development of "systems immunology" reverses this point of view and establishes a new usage of multi-parametric statistical approaches to represent the biological data by projections and "landscapes" in the N-dimensional space of considered parameters (230). Thus, the traditional description of separate repertoires for distinct cell subsets defined from a few markers is being replaced by overlapping clouds of data, setting the limits of the different classification groups (tissue of origin, infection contexts, combination of marker expression, repertoire expression. . .). Moreover, repertoire diversity technologies can now be combined to complementary approaches to decipher the complexity of lymphocyte populations, such as microwell array cell culture and high-resolution imaging (231), mass cytometry (232, 233), cellular barcoding (234), intravital imaging (235, 236), single-cell gene expression (237). In addition, high-throughput repertoire descriptions will enrich mathematical and computer models of lymphocyte repertoire diversity and dynamics such as those proposed by Mehr (238), Ciupe et al. (239), or Stirk et al. (240).

As advocated by others, the concepts developed by systems biology, such as the signatures emerging from clustering and the modularity regulating gene networks, will probably need to be adapted to the constraints of immunology data (241). However, this is probably through this kind of representation that global analysis of immune repertoires will have to be addressed (242).

The upcoming challenge is now to merge data produced through the different technological approaches available to achieve full integration of these data and make them available for interactive meta-analysis. This necessitates more than the simple juxtaposition of annotated raw data but rather requires (1) the codification and standardization of this multi-level data and (2) the integration of complexity science into immunology. Along this line, recent developments of multi-parametric flow cytometry naturally led to systematic clustering and multivariate statistical analysis approaches for searching functional signatures (2, 232, 233, 243–245).

ACKNOWLEDGMENTS

This work was supported by French state funds within the Investissements d'Avenir program (ANR-11-IDEX-0004-02; LabEx Transimmunom), the European Research Council Advanced grant (TRiPoD), the European PCRDT7 (Lifecycle program), the RNSC (ImmunoComplexiT network), CNRS (PEPS BMI), INRA and Université Pierre and Marie Curie.

REFERENCES

1. Boudinot P, Marriotti-Ferrandiz ME, Du Pasquier L, Benmansour A, Cazenave PA, Six A. New perspectives for large-scale repertoire analysis of immune receptors. *Mol Immunol* (2008) **45**:2437–45. doi:10.1016/j.molimm.2007.12.018
2. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. *Trends Immunol* (2012) **33**:323–32. doi:10.1016/j.it.2012.02.010
3. MacDonald HR, Pedrazzini T, Schneider R, Louis JA, Zinkernagel RM, Hengartner H. Intrathymic elimination of Mls^a-reactive (Vβ6$^+$) cells during neonatal tolerance induction to Mls^a-encoded antigens. *J Exp Med* (1988) **167**:2005–10. doi:10.1084/jem.167.6.2005
4. MacDonald HR, Schneider R, Lees RK, Howe RC, Acha-Orbea H, Festenstein H, et al. T-cell receptor Vβ use predicts reactivity tolerance to Mls^a-encoded antigens. *Nature* (1988) **332**:40–5. doi:10.1038/332040a0
5. Salaun J, Bandeira A, Khazaal I, Burlen-Defranoux O, Thomas-Vaslin V, Coltey M, et al. Transplantation tolerance is unrelated to superantigen-dependent deletion and anergy. *Proc Natl Acad Sci U S A* (1992) **89**:10420–4. doi:10.1073/pnas.89.21.10420
6. Faint JM, Pilling D, Akbar AN, Kitas GD, Bacon PA, Salmon M. Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression. *J Immunol Methods* (1999) **225**:53–60. doi:10.1016/S0022-1759(99)00027-7
7. Thomas-Vaslin V, Six A, Pham HP, Dansokho C, Chaara W, Gouritin B, et al. Immunodepression & Immunosuppression during aging. In: Portela MB editor. *Immunosuppression.* Rijeka: InTech open access publisher (2012). p. 125–463.
8. Lefranc MP, Lefranc G. *The Immunoglobulin FactsBook.* London: Academic Press (2001).
9. Lefranc MP, Lefranc G. *The T Cell Receptor FactsBook.* London: Academic Press (2001).
10. Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, Coutinho A. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological "homunculus" of antibodies in normal serum. *Eur J Immunol* (1993) **23**:2851–9. doi:10.1002/eji.1830231119
11. Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. *Scand J Immunol* (1994) **39**:79–87.
12. Fesel C, Coutinho A. Serum IgM repertoire reactions to MBP/CFA immunization reflect the individual status of EAE susceptibility. *J Autoimmun* (2000) **14**:319–24. doi:10.1006/jaut.2000.0373
13. Haury M, Sundblad A, Grandien A, Barreau C, Coutinho A, Nobrega A. The repertoire of serum IgM in normal mice is largely independent of external antigenic contact. *Eur J Immunol* (1997) **27**:1557–63. doi:10.1002/eji.1830270635
14. Stahl D, Lacroix-Desmazes S, Heudes D, Mouthon L, Kaveri SV, Kazatchkine MD. Altered control of self-reactive IgG by autologous IgM in patients with warm autoimmune hemolytic anemia. *Blood* (2000) **95**:328–35.
15. Stahl D, Lacroix-Desmazes S, Mouthon L, Kaveri SV, Kazatchkine MD. Analysis of human self-reactive antibody repertoires by quantitative immunoblotting. *J Immunol Methods* (2000) **240**:1–14. doi:10.1016/S0022-1759(00)00185-X
16. Costa N, Pires AE, Gabriel AM, Goulart LF, Pereira C, Leal B, et al. Broadened T-cell repertoire diversity in ivIg-treated SLE patients is also related to the individual status of regulatory T-cells. *J Clin Immunol* (2013) **33**:349–60. doi:10.1007/s10875-012-9816-7
17. Quintana FJ, Hagedorn PH, Elizur G, Merbl Y, Domany E, Cohen IR. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc Natl Acad Sci U S A* (2004) **101**:14615–21. doi:10.1073/pnas.0404848101
18. Quintana FJ, Getz G, Hed G, Domany E, Cohen IR. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bioinformatic approach to immune complexity. *J Autoimmun* (2003) **21**:65–75. doi:10.1016/S0896-8411(03)00064-7
19. Delassus S, Darche S, Kourilsky P, Cumano A. Ontogeny of the heavy chain immunoglobulin repertoire in fetal liver and bone marrow. *J Immunol* (1998) **160**:3274–80.
20. Yassai M, Gorski J. Thymocyte maturation: selection for in-frame TCRα-chain rearrangement is followed by selection for shorter TCRβ-chain complementarity-determining region 3. *J Immunol* (2000) **165**:3706–12.
21. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, et al. Quantitative and qualitative changes in V-Jα rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor α chain repertoire. *J Exp Med* (2002) **196**:1163–73. doi:10.1084/jem.20021074
22. Douek DC, McFarland RD, Keiser PH, Gage EA, Massey JM, Haynes BF, et al. Changes in thymic function with age and during the treatment of HIV infection. *Nature* (1998) **396**:690–5. doi:10.1038/25374
23. Ribeiro RM, Perelson AS. Determining thymic output quantitatively: using models to interpret experimental T-cell receptor excision circle (TREC) data. *Immunol Rev* (2007) **216**:21–34.
24. Bains I, Thiebaut R, Yates AJ, Callard R. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**:4329–36. doi:10.4049/jimmunol.0900743
25. Poulin JF, Viswanathan MN, Harris JM, Komanduri KV, Wieder E, Ringuette N, et al. Direct evidence for thymic function in adult humans. *J Exp Med* (1999) **190**:479–86. doi:10.1084/jem.190.4.479
26. Dulude G, Cheynier R, Gauchat D, Abdallah A, Kettaf N, Sékaly RP, et al. The magnitude of thymic output is genetically determined through controlled intrathymic precursor T cell proliferation. *J Immunol* (2008) **181**:7818–24.
27. Wu GE, Paige CJ. VH gene family utilization in colonies derived from B and pre-B cells detected by the RNA colony blot assay. *EMBO J* (1986) **5**:3475–81.
28. Schulze DH, Kelsoe G. Genotypic analysis of B cell colonies by in situ hybridization. Stoichiometric expression of three VH families in adult C57BL/6 and BALB/c mice. *J Exp Med* (1987) **166**:163–72. doi:10.1084/jem.166.1.163
29. Thomas-Vaslin V, Andrade L, Freitas A, Coutinho A. Clonal persistence of B lymphocytes in normal mice is determined by variable region-dependent selection. *Eur J Immunol* (1991) **21**:2239–46. doi:10.1002/eji.1830210935
30. Andrade L, Huetz F, Poncet P, Thomas-Vaslin V, Goodhardt M, Coutinho A. Biased VH gene expression in murine CD5 B cells results from age-dependent cellular selection. *Eur J Immunol* (1991) **21**:2017–23. doi:10.1002/eji.1830210908
31. Freitas AA, Lembezat MP, Coutinho A. Expression of antibody V-regions is genetically and developmentally controlled and modulated by the B lymphocyte environment. *Int Immunol* (1989) **1**:342–54. doi:10.1093/intimm/1.4.342
32. Rosenberg WMC, Moss PAH, Bell JI. Variation in human T cell receptor Vβ and Jβ repertoire: analysis using anchored polymerase reaction. *Eur J Immunol* (1992) **22**:541–9. doi:10.1002/eji.1830220237
33. Moss PAH, Rosenberg WMC, Zintzaras E, Bell JI. Characterization of the human T cell receptor α-chain repertoire and demonstration of α genetic influence on the Vα usage. *Eur J Immunol* (1993) **23**:1155–9. doi:10.1002/eji.1830230526

34. Moss PAH, Bell JI. Sequence analysis of the human αβ T-cell receptor CDR3 region. *Immunogenetics* (1995) **42**:10–8. doi:10.1007/BF00164982
35. Moss PA, Bell JI. Comparative sequence analysis of the human T cell receptor TCRA and TCRB CDR3 regions. *Hum Immunol* (1996) **48**:32–8. doi:10.1016/0198-8859(96)00084-5
36. Pannetier C, Delassus S, Darche S, Saucier C, Kourilsky P. Quantitative titration of nucleic acids by enzymatic amplification reactions run to saturation. *Nucleic Acids Res* (1993) **21**:577–83. doi:10.1093/nar/21.3.577
37. Manfras BJ, Rudert WA, Trucco M, Boehm O. Analysis of the αβ T-cell receptor repertoire by competitive and quantitative family-specific PCR with exogenous standards and high resolution fluorescence based CDR3 size imaging. *J Immunol Methods* (1997) **210**:235–49. doi:10.1016/S0022-1759(97)00197-X
38. VanderBorght A, Van der Aa A, Geusens P, Vandevyver C, Raus J, Stinissen P. Identification of overrepresented T cell receptor genes in blood and tissue biopsies by PCR-ELISA. *J Immunol Methods* (1999) **223**:47–61. doi:10.1016/S0022-1759(98)00201-4
39. Lim A, Baron V, Ferradini L, Bonneville M, Kourilsky P, Pannetier C. Combination of MHC-peptide multimer-based T cell sorting with the Immunoscope permits sensitive ex vivo quantitation and follow-up of human CD8+ T cell immune responses. *J Immunol Methods* (2002) **261**:177–94. doi:10.1016/S0022-1759(02)00004-2
40. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the αβ TCR repertoire of naive mouse splenocytes. *J Immunol* (2000) **164**:5782–7.
41. Gallard A, Foucras G, Coureau C, Guery JC. Tracking T cell clonotypes in complex T lymphocyte populations by real-time quantitative PCR using fluorogenic complementarity-determining region-3-specific probes. *J Immunol Methods* (2002) **270**:269–80. doi:10.1016/S0022-1759(02)00336-8
42. Liu DB, Callahan JP, Dau PC. Intrafamily fragment analysis of the T cell receptor β chain CDR3 region. *J Immunol Methods* (1995) **187**:139–50. doi:10.1016/0022-1759(95)00178-D
43. Mugnaini EN, Egeland T, Syversen AM, Spurkland A, Brinchmann JE. Molecular analysis of the complementarity determining region 3 of the human T cell receptor β chain. Establishment of a reference panel of CDR3 lengths from phytohaemagglutinin activated lymphocytes. *J Immunol Methods* (1999) **223**:207–16. doi:10.1016/S0022-1759(99)00004-6
44. Matsutani T, Yoshioka T, Tsuruta Y, Iwagami S, Suzuki R. Analysis of TCRAV and TCRBV repertoires in healthy individuals by microplate hybridization assay. *Hum Immunol* (1997) **56**:57–69. doi:10.1016/S0198-8859(97)00102-X
45. Matsutani T, Yoshioka T, Tsuruta Y, Iwagami S, Toyosaki-Maeda T, Horiuchi T, et al. Restricted usage of T-cell receptor α-chain variable region (TCRAV) and T-cell receptor β-chain variable region (TCRBV) repertoires after human allogeneic haematopoietic transplantation. *Br J Haematol* (2000) **109**:759–69. doi:10.1046/j.1365-2141.2000.02080.x
46. Yoshida R, Yoshioka T, Yamane S, Matsutani T, Toyosaki-Maeda T, Tsuruta Y, et al. A new method for quantitative analysis of the mouse T-cell receptor V region repertoires: comparison of repertoires among strains. *Immunogenetics* (2000) **52**:35–45. doi:10.1007/s002510000248
47. Du G, Qiu L, Shen L, Sehgal P, Shen Y, Huang D, et al. Combined megaplex TCR isolation and SMART-based real-time quantitation methods for quantitating antigen-specific T cell clones in mycobacterial infection. *J Immunol Methods* (2006) **308**:19–35. doi:10.1016/j.jim.2005.09.009
48. Ogle BM, Cascalho M, Joao C, Taylor W, West LJ, Platt JL. Direct measurement of lymphocyte receptor diversity. *Nucleic Acids Res* (2003) **31**:e139. doi:10.1093/nar/gng139
49. Joao C, Ogle BM, Gay-Rabinstein C, Platt JL, Cascalho M. B cell-dependent TCR diversification. *J Immunol* (2004) **172**:4709–16.
50. Joao C. Immunoglobulin is a highly diverse self-molecule that improves cellular diversity and function during immune reconstitution. *Med Hypotheses* (2007) **68**:158–61. doi:10.1016/j.mehy.2006.05.062
51. Bonarius HP, Baas F, Remmerswaal EB, van Lier RA, ten Berge I, Tak PP, et al. Monitoring the T-cell receptor repertoire at single-clone resolution. *PLoS One* (2006) **1**:e55. doi:10.1371/journal.pone.0000055
52. Pannetier C, Cochet M, Darche S, Casrouge A, Zöller M, Kourilsky P. The size of the CDR3 hypervariable regions of the murine T-cell receptor β chains vary as a function of the recombined germ-line segments. *Proc Natl Acad Sci U S A* (1993) **90**:4319–23. doi:10.1073/pnas.90.9.4319
53. Pannetier C, Even J, Kourilsky P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol Today* (1995) **16**:176–81. doi:10.1016/0167-5699(95)80117-0
54. Pannetier C, Levraud JP, Lim A, Even J, Kourilsky P. The immunoscope approach for the analysis of T-cell repertoires. In: Oksenberg J editor. *The Human Antigen T Cell Receptor. Selected Protocols and Applications.* Georgetown, TX: Landes RG (1997). p. 287–325.
55. Cochet M, Pannetier C, Darche S, Leclerc C, Kourilsky P. Molecular detection and *in vivo* analysis of the specific T cell response to a protein antigen. *Eur J Immunol* (1992) **22**:2639–47. doi:10.1002/eji.1830221025
56. Regnault A, Cumano A, Vassalli P, Guy-Grand D, Kourilsky P. Oligoclonal repertoire of the CD8αα and the CD8αβ TCR-α/β murine intestinal intraepithelial T lymphocytes: evidence for the random emergence of T cells. *J Exp Med* (1994) **180**:1345–58. doi:10.1084/jem.180.4.1345
57. Regnault A, Levraud JP, Lim A, Six A, Moreau C, Cumano A, et al. The expansion and selection of T cell receptor αβ intestinal intraepithelial T cell clones. *Eur J Immunol* (1996) **26**:914–21. doi:10.1002/eji.1830260429
58. Ema H, Cumano A, Kourilsky P. TCRβ repertoire development in the mouse embryo. *J Immunol* (1997) **159**:4227–32.
59. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human αβ T cell receptor diversity. *Science* (1999) **286**:958–61. doi:10.1126/science.286.5441.958
60. Bousso P, Lemaître F, Laouini D, Kanellopoulos J, Kourilsky P. The peripheral CD8 T cell repertoire is largely independent of the presence of intestinal flora. *Int Immunol* (2000) **12**:425–30. doi:10.1093/intimm/12.4.425
61. Arstila TP, Even J. Size of the αβ TCR repertoire. *Med Sci (Paris)* (2000) **16**:1257–60. doi:10.4267/10608/1566
62. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* (2001) **194**:1385–90. doi:10.1084/jem.194.9.1385
63. Fazilleau N, Cabaniols JP, Lemaitre F, Motta I, Kourilsky P, Kanellopoulos JM. Valpha and Vbeta public repertoires are highly conserved in terminal deoxynucleotidyl transferase-deficient mice. *J Immunol* (2005) **174**:345–55.
64. Cibotti R, Cabaniols JP, Pannetier C, Delarbre C, Vergnon I, Kanellopoulos JM, et al. Public and private Vβ T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice. *J Exp Med* (1994) **180**:861–72. doi:10.1084/jem.180.3.861
65. Gapin L, Fukui Y, Kanellopoulos J, Sano T, Casrouge A, Malier V, et al. Quantitative analysis of the T cell repertoire selected by a single peptide-major histocompatibility complex. *J Exp Med* (1998) **187**:1871–83. doi:10.1084/jem.187.11.1871
66. Bouneaud C, Kourilsky P, Bousso P. Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion. *Immunity* (2000) **13**:829–40. doi:10.1016/S1074-7613(00)00080-7
67. Fukui Y, Oono T, Cabaniols JP, Nakao K, Hirokawa K, Inayoshi A, et al. Diversity of T cell repertoire shaped by a single peptide ligand is critically affected by its amino acid residue at a T cell receptor contact. *Proc Natl Acad Sci U S A* (2000) **97**:13760–5. doi:10.1073/pnas.250470797
68. Musette P, Bureau JF, Gachelin G, Kourilsky P, Brahic M. T lymphocyte repertoire in Theiler's virus encephalomyelitis: the nonspecific infiltration of the central nervous system of infected SJL/J mice is associated with a selective local T cell expansion. *Eur J Immunol* (1995) **25**:1589–93. doi:10.1002/eji.1830250618
69. Sourdive DJD, Murali-Krishna K, Altman JD, Zajac AJ, Whitmire JK, Pannetier C, et al. Conserved T cell receptor repertoire in primary and memory CD8 T cell responses to an acute viral infection. *J Exp Med* (1998) **188**:71–82. doi:10.1084/jem.188.1.71
70. Musette P, Bequet D, Delarbre C, Gachelin G, Kourilsky P, Dormont D. Expansion of a recurrent Vβ5.3$^+$ T-cell population in newly diagnosed and untreated HLA-DR2 multiple sclerosis patients. *Proc Natl Acad Sci U S A* (1996) **93**:12461–6. doi:10.1073/pnas.93.22.12461
71. Fazilleau N, Delarasse C, Sweenie CH, Anderton SM, Fillatreau S, Lemonnier FA, et al. Persistence of autoreactive myelin oligodendrocyte glycoprotein (MOG)-specific T cell repertoires in MOG-expressing mice. *Eur J Immunol* (2006) **36**:533–43. doi:10.1002/eji.200535021
72. Musette P, Bachelez H, Flageul B, Delarbre C, Kourilsky P, Dubertret L, et al. Immune-mediated destruction of melanocytes in halo nevi is associated with

the local expansion of a limited number of T cell clones. *J Immunol* (1999) **162**:1789–94.
73. Douillard P, Pannetier C, Josien R, Menoret S, Kourilsky P, Soulillou JP, et al. Donor-specific blood transfusion-induced tolerance in adult rats with a dominant TCR-Vβ rearrangement in heart allografts. *J Immunol* (1996) **157**:1250–60.
74. Pilch H, Höhn H, Freitag K, Neukirch C, Necker A, Haddad P, et al. Improved assessment of T-cell receptor (TCR) VB repertoire in clinical specimens: combination of TCR-CDR3 spectratyping with flow cytometry-based TCR VB frequency analysis. *Clin Diagn Lab Immunol* (2002) **9**:257–66.
75. Messaoudi I, LeMaoult J, Guevara-Patino JA, Metzner BM, Nikolich-Zugich J. Age-related CD8 T cell clonal expansions constrict CD8 T cell repertoire and have the potential to impair immune defense. *J Exp Med* (2004) **200**:1347–58. doi:10.1084/jem.20040437
76. Hakim FT, Memon SA, Cepeda R, Jones EC, Chow CK, Kasten-Sportes C, et al. Age-dependent incidence, time course, and consequences of thymic renewal in adults. *J Clin Invest* (2005) **115**:930–9. doi:10.1172/JCI200522492
77. Messaoudi I, Warner J, Fischer M, Park B, Hill B, Mattison J, et al. Delay of T cell senescence by caloric restriction in aged long-lived nonhuman primates. *Proc Natl Acad Sci U S A* (2006) **103**:19448–53. doi:10.1073/pnas.0606661103
78. Boudinot P, Boubekeur S, Benmansour A. Rhabdovirus infection induces public and private T cell responses in teleost fish. *J Immunol* (2001) **167**:6202–9.
79. Boudinot P, Boubekeur S, Benmansour A. Primary structure and complementarity-determining region (CDR) 3 spectratyping of rainbow trout TCRβ transcripts identify ten Vβ families with Vβ6 displaying unusual CDR2 and differently spliced forms. *J Immunol* (2002) **169**:6244–52.
80. Boudinot P, Bernard D, Boubekeur S, Thoulouze MI, Bremont M, Benmansour A. The glycoprotein of a fish rhabdovirus profiles the virus-specific T-cell repertoire in rainbow trout. *J Gen Virol* (2004) **85**:3099–108. doi:10.1099/vir.0.80135-0
81. Desravines S, Hsu E. Measuring CDR3 length variability in individuals during ontogeny. *J Immunol Methods* (1994) **168**:219–25. doi:10.1016/0022-1759(94)90058-2
82. Gorski J, Yassai M, Zhu X, Kissella B, Keever C, Flomenberg N. Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 spectratyping. *J Immunol* (1994) **152**:5109–19.
83. Maslanka K, Piatek T, Gorski J, Yassai M. Molecular analysis of T cell repertoires – Spectratypes generated by multiplex polymerase chain reaction and evaluated by radioactivity or fluorescence. *Hum Immunol* (1995) **44**:28–34.
84. Lue C, Mitani Y, Crew MD, George JF, Fink LM, Schichman SA. An automated method for the analysis of T-cell receptor repertoires: rapid RT-PCR fragment length analysis of the T-cell receptor β chain complementarity-determining region 3. *Am J Clin Pathol* (1999) **111**:683–90.
85. Yamamoto K, Masuko-Hongo K, Tanaka A, Kurokawa M, Hoeger T, Nishioka K, et al. Establishment and application of a novel T cell clonality analysis using single-strand conformation polymorphism of T cell receptor messenger signals. *Hum Immunol* (1996) **48**:23–31. doi:10.1016/0198-8859(96)00080-8
86. Shiokawa S, Nishimura J, Ohshima K, Uike N, Yamamoto K. Establishment of a novel B cell clonality analysis using single-strand conformation polymorphism of immunoglobulin light chain messenger signals. *Am J Pathol* (1998) **153**:1393–400. doi:10.1016/S0002-9440(10)65726-4
87. Raaphorst FM, Gokmen E, Teale JM. Analysis of clonal diversity in mouse immunoglobulin heavy chain genes selected for size of the antigen combining site. *Immunol Invest* (1998) **27**:355–65. doi:10.3109/08820139809022709
88. Sottini A, Quirós Roldan E, Albertini A, Primi D, Imberti L. Assessment of T-cell receptor beta-chain diversity by heteroduplex analysis. *Hum Immunol* (1996) **48**:12–22. doi:10.1016/0198-8859(96)00087-0
89. Wack A, Montagna D, Dellabona P, Casorati G. An improved PCR-heteroduplex method permits high-sensitivity detection of clonal expansions in complex T cell populations. *J Immunol Methods* (1996) **196**:181–92. doi:10.1016/0022-1759(96)00114-7
90. Shen DF, Doukhan L, Kalam S, Delwart E. High-resolution analysis of T-cell receptor β-chain repertoires using DNA heteroduplex tracking: generally stable, clonal $CD8^+$ expansions in all healthy young adults. *J Immunol Methods* (1998) **215**:113–21. doi:10.1016/S0022-1759(98)00066-0
91. Wedderburn LR, Maini MK, Patel A, Beverley PCL, Woo P. Molecular fingerprinting reveals non-overlapping T cell oligoclonality between an inflamed site and peripheral blood. *Int Immunol* (1999) **11**:535–43. doi:10.1093/intimm/11.4.535
92. Bouffard P, Gagnon C, Cloutier D, MacLean SJ, Souleimani A, Nallainathan D, et al. Analysis of T cell receptor β chain expression by isoelectric focusing following gene amplification and *in vitro* translation. *J Immunol Methods* (1995) **187**:9–21. doi:10.1016/0022-1759(95)00161-3
93. Sant'Angelo DB, Lucas B, Waterbury PG, Cohen B, Brabb T, Goverman J, et al. A molecular map of T cell development. *Immunity* (1998) **9**:179–86. doi:10.1016/S1074-7613(00)80600-7
94. Correia-Neves M, Waltzinger C, Mathis D, Benoist C. The shaping of the T cell repertoire. *Immunity* (2001) **14**:21–32. doi:10.1016/S1074-7613(01)00086-3
95. Hsieh CS, Liang Y, Tyznik AJ, Self SG, Liggitt D, Rudensky AY. Recognition of the peripheral self by naturally arising CD25+ CD4+ T cell receptors. *Immunity* (2004) **21**:267–77. doi:10.1016/j.immuni.2004.07.009
96. Hsieh CS, Zheng Y, Liang Y, Fontenot JD, Rudensky AY. An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. *Nat Immunol* (2006) **7**:401–10. doi:10.1038/ni1318
97. Pacholczyk R, Ignatowicz H, Kraj P, Ignatowicz L. Origin and T cell receptor diversity of $Foxp3^+CD4^+CD25^+$ T cells. *Immunity* (2006) **25**:249–59. doi:10.1016/j.immuni.2006.05.016
98. Pacholczyk R, Kern J, Singh N, Iwashima M, Kraj P, Ignatowicz L. Nonself-antigens are the cognate specificities of $Foxp3^+$ regulatory T cells. *Immunity* (2007) **27**:493–504. doi:10.1016/j.immuni.2007.07.019
99. Wong J, Obst R, Correia-Neves M, Losyev G, Mathis D, Benoist C. Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory $CD4^+$ T cells. *J Immunol* (2007) **178**:7032–41.
100. Kang JA, Mohindru M, Kang BS, Park SH, Kim BS. Clonal expansion of infiltrating T cells in the spinal cords of SJL/J mice infected with Theiler's virus. *J Immunol* (2000) **165**:583–90.
101. Apostolou I, Cumano A, Gachelin G, Kourilsky P. Evidence for two subgroups of $CD4^-CD8^-$ NKT cells with distinct TCRαβ repertoires and differential distribution in lymphoid tissues. *J Immunol* (2000) **165**:2481–90.
102. Bernard D, Six A, Rigottier-Gois L, Messiaen S, Chilmonczyk S, Quillet E, et al. Phenotypic and functional similarity of gut intraepithelial and systemic T cells in a teleost fish. *J Immunol* (2006) **176**:3942–9.
103. Mancini S, Candéias SM, Fehling HJ, von Boehmer H, Jouvin-Marche E, Marche PN. TCR α-chain repertoire in pTα-deficient mice is diverse and developmentally regulated: implications for pre-TCR functions and TCRA gene rearrangement. *J Immunol* (1999) **163**:6053–9.
104. Halapi E, Werner A, Wahlström J, Österborg A, Jeddi-Tehrani M, Yi Q, et al. T cell repertoire in patients with multiple myeloma and monoclonal gammopathy of undetermined significance: clonal $CD8^+$ T cell expansions are found preferentially in patients with a low tumor burden. *Eur J Immunol* (1997) **27**:2245–52. doi:10.1002/eji.1830270919
105. Brawand P, Cerottini JC, MacDonald HR. Hierarchal utilization of different T-cell receptor Vβ gene segments in the $CD8^+$-T-cell response to an immunodominant Moloney leukemia virus-encoded epitope *in vivo*. *J Virol* (1999) **73**:9161–9.
106. Matsuzaki G, Takada H, Nomoto K. *Escherichia coli* infection induces only fetal thymus-derived γδ T cells at the infected site. *Eur J Immunol* (1999) **29**:3877–86. doi:10.1002/(SICI)1521-4141(199912)29:12<3877::AID-IMMU3877>3.3.CO;2-3
107. Collette A, Bagot S, Ferrandiz ME, Cazenave PA, Six A, Pied S. A profound alteration of blood TCRB repertoire allows prediction of cerebral malaria. *J Immunol* (2004) **173**:4568–75.
108. Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL, et al. A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J Immunol* (2002) **168**:3099–104.
109. Lim A, Lemercier B, Wertz X, Pottier SL, Huetz F, Kourilsky P. Many human peripheral VH5-expressing IgM+ B cells display a unique heavy-chain rearrangement. *Int Immunol* (2008) **20**:105–16. doi:10.1093/intimm/dxm125
110. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* (2009) **55**:641–58. doi:10.1373/clinchem.2008.112789
111. McGinn S, Gut IG. DNA sequencing – spanning the generations. *Nat Biotechnol* (2013) **30**:366–72. doi:10.1016/j.nbt.2012.11.012

112. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) **324**:807–10. doi:10.1126/science.1170020
113. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* (2009) **114**:4099–107. doi:10.1182/blood-2009-04-217604
114. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* (2009) **19**:1817–24. doi:10.1101/gr.092924.109
115. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A* (2010) **107**:1518–23. doi:10.1073/pnas.0913939107
116. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
117. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* (2011) **186**:4285–94. doi:10.4049/jimmunol.1003898
118. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci Transl Med* (2012) **4**:134ra63. doi:10.1126/scitranslmed.3003656
119. Föhse L, Suffner J, Suhre K, Wahl B, Lindner C, Lee CW, et al. High TCR diversity ensures optimal function and homeostasis of Foxp3$^+$ regulatory T cells. *Eur J Immunol* (2011) **41**:3101–13. doi:10.1002/eji.201141986
120. Sherwood AM, Desmarais C, Livingston RJ, Andriesen J, Haussler M, Carlson CS, et al. Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after αβ and γδ T cell commitment. *Sci Transl Med* (2011) **3**:90ra61. doi:10.1126/scitranslmed.3002536
121. Cebula A, Seweryn M, Rempala GA, Pabla SS, McIndoe RA, Denning TL, et al. Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* (2013) **497**:258–62. doi:10.1038/nature12079
122. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) **23**:1874–84. doi:10.1101/gr.154815.113
123. Srivastava SK, Robins HS. Palindromic nucleotide analysis in human T cell receptor rearrangements. *PLoS One* (2012) **7**:e52250. doi:10.1371/journal.pone.0052250
124. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* (2012) **109**:16161–6. doi:10.1073/pnas.1212755109
125. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci U S A* (2012) **109**:15865–70. doi:10.1073/pnas.1203916109
126. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8$^+$ T cell receptor repertoire. *Sci Transl Med* (2010) **2**:47ra64. doi:10.1126/scitranslmed.3001442
127. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, et al. Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* (2012) **64**:337–50. doi:10.1007/s00251-011-0595-8
128. van Heijst JW, Ceberio I, Lipuma LB, Samilo DW, Wasilewski GD, Gonzales AM, et al. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. *Nat Med* (2013) **19**:372–7. doi:10.1038/nm.3100
129. Meier J, Roberts C, Avent K, Hazlett A, Berrie J, Payne K, et al. Fractal organization of the human T cell repertoire in health and after stem cell transplantation. *Biol Blood Marrow Transplant* (2013) **19**:366–77. doi:10.1016/j.bbmt.2012.12.004
130. Ademokun A, Wu Y-C, Martin V, Mitra R, Sack U, Baxendale H, et al. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* (2011) **10**:922–30. doi:10.1111/j.1474-9726.2011.00732.x
131. Castro R, Jouneau L, Pham HP, Bouchez O, Giudicelli V, Lefranc MP, et al. Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog* (2013) **9**:e1003098. doi:10.1371/journal.ppat.1003098
132. Bousso P, Casrouge A, Altman JD, Haury M, Kanellopoulos J, Abastado JP, et al. Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* (1998) **9**:169–78. doi:10.1016/S1074-7613(00)80599-3
133. Lin MY, Welsh RM. Stability and diversity of T cell receptor repertoire usage during lymphocytic choriomeningitis virus infection of mice. *J Exp Med* (1998) **188**:1993–2005. doi:10.1084/jem.188.11.1993
134. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* (2010) **11**:31–46. doi:10.1038/nrg2626
135. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005) **437**:376–80.
136. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* (2011) **475**:348–52. doi:10.1038/nature10242
137. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* (2012) **42**:3073–83. doi:10.1002/eji.201242517
138. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol* (2013) **9**:e1003031. doi:10.1371/journal.pcbi.1003031
139. Shendure J, Aiden EL. The expanding scope of DNA sequencing. *Nat Biotechnol* (2012) **30**:1084–94. doi:10.1038/nbt.2421
140. Dash P, McClaren JL, Oguin TH III, Rothwell W, Todd B, Morris MY, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J Clin Invest* (2011) **121**:288–95. doi:10.1172/JCI44752
141. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) **31**:166–9. doi:10.1038/nbt.2492
142. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* (2013) **43**:2507–15. doi:10.1002/eji.201343453
143. Plessy C, Desbois L, Fujii T, Carninci P. Population transcriptomics with single-cell resolution: a new field made possible by microfluidics: a technology for high throughput transcript counting and data-driven definition of cell types. *Bioessays* (2013) **35**:131–40. doi:10.1002/bies.201200093
144. Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol Lett* (2012) **148**:11–22. doi:10.1016/j.imlet.2012.08.002
145. Pancer Z, Amemiya CT, Ehrhardt GR, Ceitlin J, Gartland GL, Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* (2004) **430**:174–80. doi:10.1038/nature02740
146. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* (2009) **37**:D1006–12. doi:10.1093/nar/gkn838
147. Giudicelli V, Lefranc MP. IMGT-ONTOLOGY 2012. *Front Genet* (2012) **3**:79. doi:10.3389/fgene.2012.00079
148. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP. IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res* (2012) **8**:26. doi:10.1007/978-1-61779-842-9_32
149. Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* (2005) **309**:1874–8. doi:10.1126/science.1116887
150. Du Pasquier L. Insects diversify one molecule to serve two systems. *Science* (2005) **309**:1826–7. doi:10.1126/science.1118828
151. Zhang SM, Adema CM, Kepler TB, Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science* (2004) **305**:251–4. doi:10.1126/science.1088069
152. Philipp EER, Kraemer L, Melzner F, Poustka AJ, Thieme S, Findeisen U, et al. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the Lophotrochozoan *Mytilus edulis*. *PLoS One* (2012) **7**:e33091. doi:10.1371/journal.pone.0033091

153. Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, et al. Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. *BMC Genomics* (2007) **8**:124. doi:10.1186/1471-2164-8-124
154. Matsushima N, Miyashita H, Mikami T, Kuroki Y. A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. *BMC Microbiol* (2010) **10**:235. doi:10.1186/1471-2180-10-235
155. Kaas Q, Ehrenmann F, Lefranc MP. IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? *Brief Funct Genomic Proteomic* (2007) **6**:253–64. doi:10.1093/bfgp/elm032
156. Bomberger C, Singh-Jairam M, Rodey G, Guerriero A, Yeager AM, Fleming WH, et al. Lymphoid reconstitution after autologous PBSC transplantation with FACS-sorted CD34+ hematopoietic progenitors. *Blood* (1998) **91**:2588–600.
157. Gorochov G, Neumann AU, Kereveur A, Parizot C, Li TS, Katlama C, et al. Perturbation of $CD4^+$ and $CD8^+$ T-cell repertoires during progression to AIDS and regulation of the $CD4^+$ repertoire during antiviral therapy. *Nat Med* (1998) **4**:215–21. doi:10.1038/nm0298-215
158. Wu CJ, Chillemi A, Alyea EP, Orsini E, Neuberg D, Soiffer RJ, et al. Reconstitution of T-cell receptor repertoire diversity following T-cell depleted allogeneic bone marrow transplantation is related to hematopoietic chimerism. *Blood* (2000) **95**:352–9.
159. Hori S, Collette A, Demengeot J, Stewart J. A new statistical method for quantitative analyses: application to the precise quantification of T cell receptor repertoires. *J Immunol Methods* (2002) **268**:159–70. doi:10.1016/S0022-1759(02)00187-4
160. Collette A, Six A. ISEApeaks: an excel platform for GeneScan and Immunoscope data retrieval, management and analysis. *Bioinformatics* (2002) **18**:329–30. doi:10.1093/bioinformatics/18.2.329
161. Guillet M, Brouard S, Gagne K, Sebille F, Cuturi MC, Delsuc MA, et al. Different qualitative and quantitative regulation of Vβ TCR transcripts during early acute allograft rejection and tolerance induction. *J Immunol* (2002) **168**:5088–95.
162. Peggs KS, Verfuerth S, D'Sa S, Yong K, Mackinnon S. Assessing diversity: immune reconstitution and T-cell receptor BV spectratype analysis following stem cell transplantation. *Br J Haematol* (2003) **120**:154–65. doi:10.1046/j.1365-2141.2003.04036.x
163. Long SA, Khalili J, Ashe J, Berenson R, Ferrand C, Bonyhadi M. Standardized analysis for the quantification of Vbeta CDR3 T-cell receptor diversity. *J Immunol Methods* (2006) **317**:100–13. doi:10.1016/j.jim.2006.09.015
164. Miqueu P, Guillet M, Degauque N, Dore JC, Soulillou JP, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol* (2007) **44**:1057–64. doi:10.1016/j.molimm.2006.06.026
165. Guillet M, Sebille F, Soulillou JP. TCR usage in naive and committed alloreactive cells: implications for the understanding of TCR biases in transplantation. *Curr Opin Immunol* (2001) **13**:566–71. doi:10.1016/S0952-7915(00)00260-0
166. Collette A, Cazenave PA, Pied S, Six A. New methods and software tools for high throughput CDR3 spectratyping. Application to T lymphocyte repertoire modifications during experimental malaria. *J Immunol Methods* (2003) **278**:105–16. doi:10.1016/S0022-1759(03)00225-4
167. Sassi A, Largueche-Darwaz B, Collette A, Six A, Laouini D, Cazenave PA, et al. Mechanisms of the natural reactivity of lymphocytes from noninfected individuals to membrane-associated *Leishmania infantum* antigens. *J Immunol* (2005) **174**:3598–607.
168. Castro R, Takizawa F, Chaara W, Lunazzi A, Dang TH, Koellner B, et al. Contrasted TCRβ diversity of $CD8^+$ and $CD8^-$ T cells in rainbow trout. *PLoS One* (2013) **8**:e60175. doi:10.1371/journal.pone.0060175
169. Kepler TB, He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML. Statistical analysis of antigen receptor spectratype data. *Bioinformatics* (2005) **21**:3394–400. doi:10.1093/bioinformatics/bti539
170. He M, Tomfohr JK, Devlin BH, Sarzotti M, Markert ML, Kepler TB. SpA: web-accessible spectratype analysis: data management, statistical analysis and visualization. *Bioinformatics* (2005) **21**:3697–9. doi:10.1093/bioinformatics/bti600
171. Liu C, He M, Rooney B, Kepler TB, Chao NJ. Longitudinal analysis of T-cell receptor variable beta chain repertoire in patients with acute graft-versus-host disease after allogeneic stem cell transplantation. *Biol Blood Marrow Transplant* (2006) **12**:335–45. doi:10.1016/j.bbmt.2005.09.019
172. Lefranc MP. From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* (2011) **2011**:627–32. doi:10.1101/pdb.ip84
173. Lefranc MP. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc* (2011) **2011**:614–26. doi:10.1101/pdb.ip84
174. Lefranc MP. From IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes. *Cold Spring Harb Protoc* (2011) **2011**:604–13. doi:10.1101/pdb.ip84
175. Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) **2011**:633–42. doi:10.1101/pdb.ip85
176. Giudicelli V, Chaume D, Lefranc MP. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) **32**:W435–40. doi:10.1093/nar/gkh412
177. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* (2008) **36**:W503–8. doi:10.1093/nar/gkn316
178. Gaëta BA, Malming HR, Jackson KJL, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* (2007) **23**:1580–7. doi:10.1093/bioinformatics/btm147
179. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) **3**:176. doi:10.3389/fimmu.2012.00176
180. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) **41**:W34–40. doi:10.1093/nar/gkt382
181. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) **29**:542–50. doi:10.1093/bioinformatics/btt004
182. Pham HP, Manuel M, Petit N, Klatzmann D, Cohen-Kaminsky S, Six A, et al. Half of the T-cell repertoire combinatorial diversity is genetically determined in humans and humanized mice. *Eur J Immunol* (2012) **42**:760–70. doi:10.1002/eji.201141798
183. Eisenstein M. Personalized, sequencing-based immune profiling spurs startups. *Nat Biotechnol* (2013) **31**:184–6. doi:10.1038/nbt0313-184b
184. Stahl D, Lacroix-Desmazes S, Barreau C, Sibrowski W, Kazatchkine MD, Kaveri SV. Altered antibody repertoires of plasma IgM and IgG toward nonself antigens in patients with warm autoimmune hemolytic anemia. *Hum Immunol* (2001) **62**:348–61. doi:10.1016/S0198-8859(01)00225-7
185. Magurran AE. *Measuring Biological Diversity*. Oxford: Wiley-Blackwell (2004).
186. Colwell RK. *EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples. [Version 9]. User's Guide and application* (2013). Available from: http://purl.oclc.org/estimates
187. Wu TT, Kabat EA. An analysis of the sequence of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) **132**:211–50. doi:10.1084/jem.132.2.211
188. Jores R, Alzari PM, Meo T. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. *Proc Natl Acad Sci U S A* (1990) **87**:9138–42. doi:10.1073/pnas.87.23.9138
189. Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, et al. A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol Immunol* (1997) **34**:1067–82. doi:10.1016/S0161-5890(97)00130-2
190. Thomas PG, Handel A, Doherty PC, La Gruta NL. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naïve and immune T-cell populations. *Proc Natl Acad Sci U S A* (2013) **110**:1839–44. doi:10.1073/pnas.1222149110
191. Li S, Lefranc MP, Miles J, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) **4**:2333. doi:10.1038/ncomms3333
192. Conrad JA, Ramalingam RK, Duncan CB, Smith RM, Wei J, Barnett L, et al. Antiretroviral therapy reduces the magnitude and T cell receptor repertoire

diversity of HIV-specific T cell responses without changing T cell clonotype dominance. *J Virol* (2012) **86**:4213–21. doi:10.1128/JVI.06000-11
193. Koning D, Costa AI, Hoof I, Miles JJ, Nanlohy NM, Ladell K, et al. CD8$^+$ TCR repertoire formation is guided primarily by the peptide component of the antigenic complex. *J Immunol* (2013) **190**:931–9. doi:10.4049/jimmunol.1202466
194. Johnson PLF, Yates AJ, Goronzy JJ, Antia R. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. *Proc Natl Acad Sci U S A* (2012) **109**:21432–7. doi:10.1073/pnas.1209283110
195. Perelson AS, Weisbuch G. Immunology for physicist. *Rev Mod Phys* (1997) **69**:1219–67. doi:10.1103/RevModPhys.69.1219
196. Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* (1979) **81**:645–70. doi:10.1016/0022-5193(79)90275-3
197. Percus JK, Percus OE, Perelson AS. Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proc Natl Acad Sci U S A* (1993) **90**:1691–5. doi:10.1073/pnas.90.5.1691
198. Bergstrom CT, Antia R. How do adaptive immune systems control pathogens while avoiding autoimmunity? *Trends Ecol Evol* (2006) **21**:22–8. doi:10.1016/j.tree.2005.11.008
199. Perelson AS. Modelling viral and immune system dynamics. *Nat Rev Immunol* (2002) **2**:28–36. doi:10.1038/nri700
200. Antia R, Ganusov VV, Ahmed R. The role of models in understanding CD8$^+$ T-cell memory. *Nat Rev Immunol* (2005) **5**:101–11. doi:10.1038/nri1550
201. Thomas-Vaslin V, Six A, Bellier B, Klatzmann D. Lymphocytes dynamics repertoires, modeling. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H editors. *Encyclopedia of Systems Biology*. Heidelberg: Springer Verlag (2013). p. 1149–52. doi:10.1007/978-1-4419-9863-7_96
202. De Boer RJ, Homann D, Perelson AS. Different dynamics of CD4$^+$ and CD8$^+$ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J Immunol* (2003) **171**:3928–35.
203. Verkoczy LK, Martensson AS, Nemazee D. The scope of receptor editing and its association with autoimmunity. *Curr Opin Immunol* (2004) **16**:808–14. doi:10.1016/j.coi.2004.09.017
204. Wucherpfennig KW, Allen PM, Celada F, Cohen IR, De BR, Garcia KC, et al. Polyspecificity of T cell and B cell receptor recognition. *Semin Immunol* (2007) **19**:216–24. doi:10.1016/j.smim.2007.02.012
205. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser IDC. Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol* (2011) **29**:527–85. doi:10.1146/annurev-immunol-030409-101317
206. Emonet T, Altan-Bonnet G. Systems immunology: a primer for biophysicists. In: Egelman E editor. *Comprehensive Biophysics*. New York: Academic Press (2012). p. 389–413.
207. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107
208. Martins VC, Ruggiero E, Schlenner SM, Madan V, Schmidt M, Fink PJ, et al. Thymus-autonomous T cell development in the absence of progenitor import. *J Exp Med* (2012) **209**:1409–17. doi:10.1084/jem.20120846
209. Farmer JD, Packard NH, Perelson AS. The immune system, adaptation and machine learning. *Physica D* (1986) **22**:187–204. doi:10.1016/0167-2789(86)90240-X
210. De Boer RJ, Perelson AS. Size and connectivity as emergent properties of a developing immune network. *J Theor Biol* (1991) **149**:381–424. doi:10.1016/S0022-5193(05)80313-3
211. Celada F, Seiden PE. A computer model of cellular interactions in the immune system. *Immunol Today* (1992) **13**:56–62. doi:10.1016/0167-5699(92)90135-T
212. Goldstein B, Faeder JR, Hlavacek WS. Mathematical and computational models of immune-receptor signalling. *Nat Rev Immunol* (2004) **4**:445–56. doi:10.1038/nri1374
213. Chao A, Chazdon RL, Colwell RK, Shen TJ. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* (2005) **8**:148–59. doi:10.1111/j.1461-0248.2004.00707.x
214. Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci U S A* (2008) **105**:16671–6. doi:10.1073/pnas.0808081105
215. Kosmrlj A, Chakraborty AK, Kardar M, Shakhnovich EI. Thymic selection of T-cell receptors as an extreme value problem. *Phys Rev Lett* (2009) **103**:068103. doi:10.1103/PhysRevLett.103.068103
216. Verhagen J, Genolet R, Britton GJ, Stevenson BJ, Sabatos-Peyton CA, Dyson J, et al. CTLA-4 controls the thymic development of both conventional and regulatory T cells through modulation of the TCR repertoire. *Proc Natl Acad Sci U S A* (2013) **110**:E221–30. doi:10.1073/pnas.1208573110
217. Mora T, Walczak AM, Bialek W, Callan CG. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci U S A* (2010) **107**:5405–10. doi:10.1073/pnas.1001705107
218. Baum PD, Venturi V, Price DA. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* (2012) **42**:2834–9. doi:10.1002/eji.201242999
219. Robins H, Desmarais C, Matthis J, Livingston R, Andriesen J, Reijonen H, et al. Ultra-sensitive detection of rare T cell clones. *J Immunol Methods* (2012) **375**:14–9. doi:10.1016/j.jim.2011.09.001
220. Zipf GK. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley (1949).
221. Sepulveda N, Paulino CD, Carneiro J. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J Immunol Methods* (2009) **353**:124–37. doi:10.1016/j.jim.2009.11.009
222. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. *J Theor Biol* (2011) **269**:1–15. doi:10.1016/j.jtbi.2010.10.001
223. Ben-Hamo R, Efroni S. The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC Syst Biol* (2011) **5**:27. doi:10.1186/1752-0509-5-27
224. Bleakley K, Lefranc MP, Biau G. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. *BMC Bioinformatics* (2008) **9**:408. doi:10.1186/1471-2105-9-408
225. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol* (2003) **171**:4639–49.
226. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, et al. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol* (2009) **183**:7314–25. doi:10.4049/jimmunol.0902452
227. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) **39**:W499–504. doi:10.1093/nar/gkr413
228. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) **40**:e134. doi:10.1093/nar/gks457
229. Rock EP, Sibbald PR, Davis MM, Chien Y. CDR3 length in antigen-specific immune receptors. *J Exp Med* (1994) **179**:323–8. doi:10.1084/jem.179.1.323
230. Hyatt G, Melamed R, Park R, Seguritan R, Laplace C, Poirot L, et al. Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol* (2006) **7**:686–91. doi:10.1038/ni0706-686
231. Han Q, Bagheri N, Bradshaw EM, Hafler DA, Lauffenburger DA, Love JC. Polyfunctional responses by human T cells result from sequential release of cytokines. *Proc Natl Acad Sci U S A* (2012) **109**:1607–12. doi:10.1073/pnas.1117194109
232. Bendall SC, Simonds EF, Qiu P, Amir E, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* (2011) **332**:687–96. doi:10.1126/science.1198704
233. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8$^+$ T cell phenotypes. *Immunity* (2012) **36**:142–52. doi:10.1016/j.immuni.2012.01.002
234. Schepers K, Swart E, van Heijst JW, Gerlach C, Castrucci M, Sie D, et al. Dissecting T cell lineage relationships by cellular barcoding. *J Exp Med* (2008) **205**:2309–18. doi:10.1084/jem.20072462
235. Sumen C, Mempel TR, Mazo IB, von Andrian UH. Intravital microscopy: visualizing immunity in context. *Immunity* (2004) **21**:315–29. doi:10.1016/j.immuni.2004.08.006
236. Marangoni F, Murooka TT, Manzo T, Kim EY, Carrizosa E, Elpek NM, et al. The transcription factor NFAT exhibits signal memory during serial T cell interactions with antigen-presenting cells. *Immunity* (2013) **38**:237–49. doi:10.1016/j.immuni.2012.09.012

237. Flatz L, Roychoudhuri R, Honda M, Filali-Mouhim A, Goulet JP, Kettaf N, et al. Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proc Natl Acad Sci U S A* (2011) **108**:5724–9. doi:10.1073/pnas.1013084108
238. Mehr R. Modeling and analysis of the meta-population dynamics of lymphocyte repertoires. *J Comput Appl Math* (2005) **184**:223–41. doi:10.1016/j.cam.2004.07.033
239. Ciupe SM, Devlin BH, Markert ML, Kepler TB. The dynamics of T-cell receptor repertoire diversity following thymus transplantation for DiGeorge anomaly. *PLoS Comput Biol* (2009) **5**:e1000396. doi:10.1371/journal.pcbi.1000396
240. Stirk ER, Molina-Paris C, van den Berg HA. Stochastic niche structure and diversity maintenance in the T cell repertoire. *J Theor Biol* (2008) **255**:237–49. doi:10.1016/j.jtbi.2008.07.017
241. Benoist C, Germain RN, Mathis D. A plaidoyer for 'systems immunology.' *Immunol Rev* (2006) **210**:229–34. doi:10.1111/j.0105-2896.2006.00374.x
242. Cohen IR. Autoantibody repertoires, natural biomarkers, and system controllers. *Trends Immunol* (2013). doi:10.1016/j.it.2013.05.003
243. Petrausch U, Haley D, Miller W, Floyd K, Urba WJ, Walker E. Polychromatic flow cytometry: a rapid method for the reduction and analysis of complex multiparameter data. *Cytometry* (2006) **69A**:1162–73. doi:10.1002/cyto.a.20342
244. Hofmann M, Zerwes HG. Identification of organ-specific T cell populations by analysis of multiparameter flow cytometry data using DNA-chip analysis software. *Cytometry A* (2006) **69**:533–40.
245. Lugli E, Pinti M, Troiano L, Nasi M, Patsekin V, Robinson JP, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. *Cytometry A* (2007) **71A**:334–44. doi:10.1002/cyto.a.20387

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2013; accepted: 12 November 2013; published online: 27 November 2013.
*Citation: Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham H-P, Lefranc M-P, Mora T, Thomas-Vaslin V, Walczak AM and Boudinot P (2013) The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. Front. Immunol. **4**:413. doi: 10.3389/fimmu.2013.00413*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Six, Mariotti-Ferrandiz, Chaara, Magadan, Pham, Lefranc, Mora, Thomas-Vaslin, Walczak and Boudinot. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

TECHNOLOGY REPORT
published: 23 December 2013
doi: 10.3389/fimmu.2013.00456

Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling

Ilgar Z. Mamedov[1,2†], Olga V. Britanova[1†], Ivan V. Zvyagin[1,2], Maria A. Turchaninova[1], Dmitriy A. Bolotin[1], Ekaterina V. Putintseva[1], Yuriy B. Lebedev[1] and Dmitriy M. Chudakov[1,2]*

[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia*
[2] *CEITEC, Masaryk University, Brno, Czech Republic*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Magadan Mompo Susana, Institut National de la Recherche Agronomique, France
Kathrin Kalies, Institute of Anatomy, Germany

***Correspondence:**
Dmitriy M. Chudakov, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Miklukho-Maklaya 16/10, 117997 Moscow, Russia
e-mail: chudakovdm@mail.ru

† *Ilgar Z. Mamedov and Olga V. Britanova have contributed equally to this work.*

High-throughput sequencing has the power to reveal the nature of adaptive immunity as represented by the full complexity of T-cell receptor (TCR) and antibody (IG) repertoires, but is at present severely compromised by the quantitative bias, bottlenecks, and accumulated errors that inevitably occur in the course of library preparation and sequencing. Here we report an optimized protocol for the unbiased preparation of TCR and IG cDNA libraries for high-throughput sequencing, starting from thousands or millions of live cells in an investigated sample. Critical points to control are revealed, along with tips that allow researchers to minimize quantitative bias, accumulated errors, and cross-sample contamination at each stage, and to enhance the subsequent bioinformatic analysis. The protocol is simple, reliable, and can be performed in 1–2 days.

Keywords: TCR repertoires, BCR repertoires, NGS applications, cDNA libraries, MiTCR, IG repertoires, T-cell receptor, T-cell receptor repertoire

INTRODUCTION

Next generation sequencing (NGS) technologies opened a breathtaking opportunity to perform deep analysis and comparative studies of the T-cell receptor (TCR) and antibody (IG) repertoires of the human donors and model animals, as well as of the various sorted, separated, or cultured lymphocyte subsets of interest (1–13). Still, rational NGS-analysis of such immune repertoires is critically dependent on the library preparation protocols, starting from a lymphocytes/PBMC sample and ending with the amplification of individual TCR/IG segment encoding molecules on the solid phase of a sequencing machine. Multiple sampling bottlenecks, PCR biases, and cross-contamination at different stages lie in wait to trick a researcher on his way to get the deep, clear, and congruent data.

While studying autoimmunity and hematopoietic stem cell transplantation therapy (10, 14–17), we have optimized cDNA-based protocol that allows unbiased pre-sequencing amplification of the human and murine, alpha- and beta-TCR, as well as IG heavy chain gene libraries. The protocol employs a specific oligonucleotide to prime cDNA synthesis, and template switching effect to form a universal 5′-adapter and to introduce sample barcode at the very first stage of library preparation. Subsequent two-step PCR amplification is performed with universal pairs of primers for the whole library using step-out plus PCR-suppression effect (18) on the 5′-end and nested PCR (19) on the 3′-end of the library (16).

This approach allows efficient and unbiased amplification of millions of the TCR/IG mRNA molecules in only 27–30 (21–24 considering dilution factor, see below) PCR cycles, thus providing sufficient starting material for the deep NGS-analysis of complex lymphocyte samples. Current protocol is optimal for the sequencing on Illumina MiSeq/HiSeq platforms and Roche 454 platforms.

Here we report the upgraded and tested protocol in a ready-to-use format with the technical details required for the method to be easily and uniformly reproduced in any laboratory.

ADVANTAGES OF cDNA LIBRARIES AND 5′-TEMPLATE SWITCH

Starting with cDNA synthesis using 5′-template switching (16, 20, 21) has at least two decisive advantages in comparison with the genomic DNA-based approaches (2, 12).

First, the whole diversity of variable chains (up to approximately 100 different V gene segment variants[1], can be amplified using just a pair (for TCRs) or a simple multiplex set (for IGs) of oligonucleotides, specific to the template switch adapter on the 5′-end and to the constant gene segments on the 3′-end of the library (**Figure 1**).

In contrast, the approaches starting with the genomic DNA require multiplex primer sets to be used both at the 5′ V gene segments' end, and at the 3′ introns/J-segments end of the library (2). Moreover, a subsequent nested PCR amplification, which requires

[1] http://www.imgt.org/

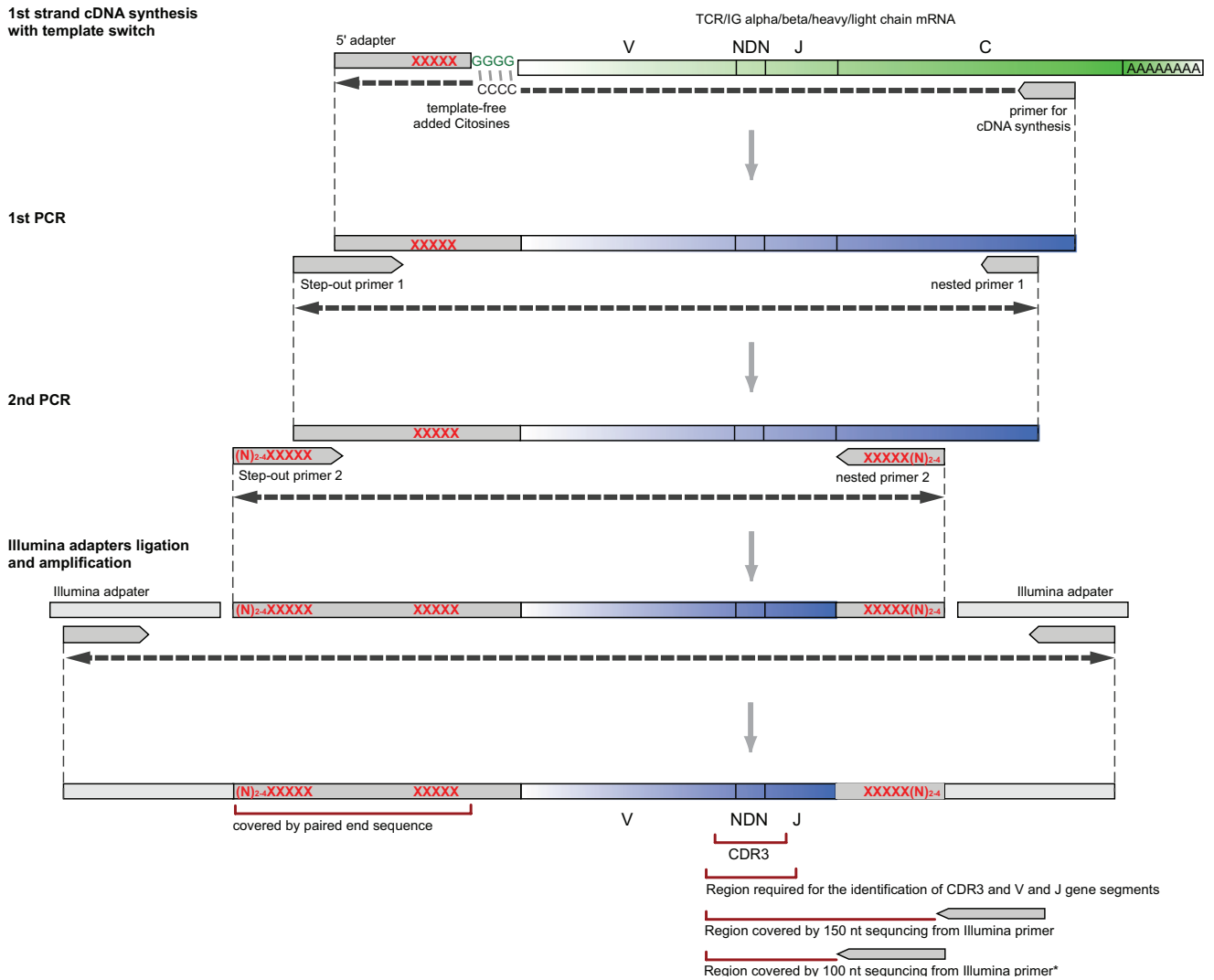


FIGURE 1 | Flow-chart of the library preparation protocol from RNA and to NGS-ready PCR product. XXXXX: optional sample barcodes (see Sample Barcoding in Appendix for details and Supplementary Material for barcodes). *For TCR alpha/beta profiling with 100 nt sequencing length, multiplexed J-segment-specific primers should be used as a reverse primer in the second PCR amplification step as described in section "Next Generation Sequencing Options."

another set of multiplex primers, can be necessary to obtain pure TCR or IG library from genomic DNA. Multiplexing inevitably leads to dramatic bias in relative efficiency of amplification of different variable segments and thus to the loss of quantitative information, and complete loss of some of the rare clonotypes (10, 16, 22, 23).

Second, abundant copies of mRNAs encoding TCR or IG chains comprise an essential portion of the total lymphocyte RNA. This practically results in an efficient amplification of a deep library starting from 10^6 mRNA molecules from a 3 μg of total RNA sample purified from three million PBMC cells (10). cDNA synthesis reaction can be performed in a volume of 10–15 μl in a single PCR tube (see Protocol), allowing multiple parallel experiments to be carried out.

In contrast, amplification of the TCR/IG library starting from 15 μg of genomic DNA of the same three million PBMC sample requires PCR to be carried out in larger volumes (since no more than 0.5 μg of genomic DNA can be taken for a 50 μl PCR reaction), and still does not provide comparable PCR efficiency, i.e., essential portion of the original sample diversity is lost due to the stochastic character of PCR, inevitably missing rare molecules.

LIMITATIONS OF THE USE OF cDNA LIBRARIES AND 5′-TEMPLATE SWITCH

We have recently demonstrated that cDNA-based template switching protocol is highly quantitative at the ensemble level – the level of relative TRBV gene segments' frequencies (10). Indeed, PCR bias is minimized and the whole approach is quite quantitative

in respect of relative abundance of mRNA molecules at start and sequencing reads at the end of analysis pipeline. However, it should be noted that individual T-cell or B-cell clones can potentially be characterized by higher or lower expression levels of TCR or IG mRNA (24, 25). This limitation should be kept in mind when using NGS data for the estimation of particular lymphocyte clones' relative abundance.

It is generally important that the cells being analyzed "feel fine" and contain a sufficient amount of TCR/IG mRNA. Therefore, it is preferable to purify total RNA from a freshly isolated cell sample for the native analysis. For the frozen samples, overnight incubation of thawed cells in presence of IL2 (Roche, 15 U/ml) leads to at least twofold increase of TCR genes RNA expression levels (our unpublished observations).

Differences in the efficiency of reverse transcription and template switching may lead to a different number of cDNA molecules read per T- or B-cell. Therefore, it is important to use the same reverse transcriptase and 5′-template switch adapter and carry out all the procedures in identical experimental conditions to obtain results that can be further accurately compared at the deep level (e.g., in an analysis of relative diversity of naïve T cells or a PBMC sample, etc.).

EXPERIMENTAL DESIGN: CELLS, NUMBERS, AND BOTTLENECKS

The desirable depth of TCR or IG repertoire analysis depends on the particular experimental questions raised. For example, application of the current protocol for the deep analysis of a PBMC sample containing 10^6 T cells will provide quantitative data on those TCR clonotypes that constitute at least 0.01–0.1% of all T cells in a sample (100–1000 T cells) (10). The majority (>95%) of TCR clonotypes constituting at least 0.001% (at least 10 T cells) will be sequenced, while approximately 20–40% of TCR clonotypes represented by a single T cell in a sample may be lost (estimated according to our quantitative experiments, depends on the reverse transcriptase used). Preferably, all the synthesized cDNA should be used for the first PCR amplification step. Second PCR should result in sufficient amount of target PCR product in a reasonable number of amplification cycles (see Protocol). The desirable number of output CDR3-containing high quality sequencing reads is at least 2×10^6 per sample (see Protocol and Expected Results).

Much smaller bottleneck limits should be quite sufficient for the majority of the experimental tasks concerning more specific subpopulations of lymphocytes characterized by lower diversity [such as sorted antigen-specific T cells (26) or B cells (27)]. For example, 10,000 lymphocytes, 10 ng high quality total RNA, no more than 21 first PCR cycles, no more than 20 s PCR cycles (see Protocol and Expected Results), and at least 30,000 CDR3-containing sequencing reads (ideally 100,000 reads to achieve over-sequencing) per sample may be sufficient to identify most TCR/IG clonotypes in a low-complexity sample. It is preferable to use reverse transcriptase with high 5′-template switching efficiency (e.g., SMARTScribe, Clontech) when small cell samples/RNA amounts are analyzed.

EXPERIMENTAL DESIGN: SAMPLE BARCODES, MULTIPLEXED SEQUENCING, CROSS-CONTAMINATION

Since as few as 30,000 sequencing reads per sample may be sufficient for many experimental tasks in immune repertoire's profiling, and, for example, paired end 150 bp Illumina MiSeq run can produce more than five million good quality TCR/IG CDR3 reads, a researcher may be often interested in sequencing multiple samples in a single run. At the same time, ligating Illumina sample barcodes to 10 or more samples is rather expensive and laborious. Our design suggests that sample barcodes can be introduced within the 5′-template switch adapter during cDNA synthesis and/or second PCR amplification steps (see **Figure 1**). Samples with the barcodes inside can be then combined in equal (or unequal, if it is desirable to get more reads for some samples) proportions, and Illumina adapters can be ligated to the resulting pooled PCR library of approximately 500–600 bp length (see Protocol and Sample Barcoding in Appendix).

Sample barcodes on both ends of the library allow to eliminate most cross-contaminations between the samples sequenced in the same run/lane that may occur during the amplification of the combined sample after adapters' ligation, and potentially in course of bridge amplification on the solid phase of the sequencing machine.

To avoid contamination on the earlier stages of pre-sequencing library preparation, all procedures, including: RNA purification, cDNA synthesis, first and second PCR preparation – should be performed in separate clean PCR boxes.

PROTOCOL

PREPARING STARTING MATERIAL – TOTAL RNA

1. Use standard Trizol (Invitrogen) or QIAzol (QIAGEN), or other analogous protocol for RNA isolation. Alternatively, use RNeasy kit (QIAGEN), or other column-based RNA isolation method. Depending on the starting material, consider the following RNA purification procedures:
   A. For small amount of whole blood (less than 100 μl) use 1 ml of Trizol or specific RNA isolation kits (for example, QIAamp RNA Blood Kit, QIAGEN).
   B. For large amount of whole blood, preferably perform preliminary PBMC separation using standard procedures (Ficoll density gradient separation) and proceed to C.
   C. For large amount of white blood cells, use 1 ml of Trizol (per up to 10^7 cells). If using column-based RNA isolation method for the large amount of cells, DNase treatment is necessary (according to a manufacturer protocol) since large amounts of genomic DNA significantly affect cDNA synthesis.
   D. For small amount of cells (below 100,000 live cells, for example, sorted or bead-separated T or B cells), preferably perform isolation of total RNA shortly after cell acquisition, in order to minimize loss of live cells and mRNA. When using Trizol protocol, add a co-precipitant (e.g., Pellet Paint, Millipore) to the aqueous phase before adding isopropyl alcohol. It is highly desirable that the precipitant forms a single well-defined spot. This provides confidence that some portion of the material will not be washed off by EtOH. Do not discard EtOH used to wash the sample until you are convinced that library preparation has been performed successfully, since some portion of RNA can remain in EtOH.

All the cell/RNA isolation, cDNA synthesis and first PCR preparation steps should be carried out in a clean DNA/RNAase free

room or a PCR box with no contact with any TCR-containing PCR products to prevent contamination. Standard RNA samples handling precautions should be used (gloves, labcoats, filtered tips, and certified RNAase free reagents) to avoid RNA degradation.

Time: 1–2 h.
Pause: RNA can be stored in 70% ethanol at −70°C for at least a year.

cDNA SYNTHESIS AND TEMPLATE SWITCH

2. Mix the following in a final volume of 4 μl in a sterile thin-walled reaction tube (mix1).

Component	Amount, μl	Final concentration*
RNA	1–3	Maximum 2 μg
cDNA synthesis primer(s) (20 μM)**	0.5–1.5 (0.5 each)	1 μM for each primer
mQ	0–2.5	

**Final concentration/amount in 10 μl after adding mix2 (see Step 5).*
***See **Table 1** for primers used. Simultaneous synthesis of TCR alpha and beta cDNA is possible (tested for both human and mouse) in case if limited starting material is available. Simultaneous synthesis of IgA, IgM, and IgG heavy chains cDNA is also possible (tested for human).*

Put no more than 1.5–2 μg of total RNA per 10 μl of final reaction volume. For the extra-deep profiling use proportional volume to obtain cDNA from desired amount of starting RNA.

3. Place the reaction tube(s) into a thermal cycler and incubate for 4 min at 70°C and then for 2 min at 42°C to anneal synthesis primer(s).
4. While incubating, mix the following in a separate tube in a final volume of 6 μl (mix2).

Component	Amount, μl	Final concentration*
First strand buffer (5×, Evrogen or Clontech)	2	1×
DTT (20 μM)	1	2 μM
5′-template switch adapter (10 μM)	1	1 μM
dNTP solution (10 mM each)	1	1 mM each
Mint reverse transcriptase (10×, Evrogen) or SMARTScribe reverse transcriptase (10×, Clontech)	1	1×

**Final concentration in 10 μl after adding mix 2.*

5. Add mix2 to mix1 and mix by pipetting, incubate 40–60 min at 42°C.
   Reverse transcriptases are heat sensitive. Allow the mixture to chill to 42°C after first step denaturation at least for 2 min as described.
   Reverse transcriptases are not equal in their 5′-template switching activity. We have extensive experience with Mint and SMARTScribe reverse transcriptases that provide reliable 5′-template switching.
6. (Optional, for Mint Reverse transcriptase only, to enhance template switching activity) Add 5 μl of IP solution (Evrogen) and incubate at 42°C for additional 1 h.
7. (Optional, see Unique Molecular Identifiers in Appendix) Add 1 μl of Uracyl DNA glycosylase (5 U/μl, New England Biolabs) and incubate 1 h at 37°C.

Time: 2–3 h.
Pause: although cDNA is generally stable, we prefer not to store cDNA longer than several hours at +4°C for the deep profiling experiments. Freezing small amounts of cDNA is undesirable.

FIRST PCR AMPLIFICATION

8. In a sterile thin-walled tube(s) mix the following in a final volume of 25 μl.

Component	Amount, μl	Final concentration
First strand cDNA	1	
Tersus buffer (10×, Evrogen)	2, 5	1×
dNTP (2.5 mM each)	1, 5	0.15 mM each
Primer smart20 (10 μM)	1	0.4 μM
Reverse primer(s) (10 μM)*	1–3 (1 each)	0.4 μM (each)
Tersus polymerase mix (50×, Evrogen)	0.5	1×
mQ	17.5–15.5	

See **Table 1 for primers used. Simultaneous amplification of TCR alpha and beta cDNA is possible (tested for both human and mouse) in case if limited starting material is available. Simultaneous amplification of IgA, IgM, and IgG heavy chains cDNA is also possible (tested for human).*

Put no more than 1 μl of cDNA from the synthesis reaction per 25 μl PCR reaction volume. For the deep profiling, use proportional number of tubes to amplify all the cDNA obtained.
Polymerase with high fidelity and processivity should be used for amplification.

9. Carry out 18 (when starting from large amount of cells) or 21 (when starting from small amount of cells) cycles of amplification using the following program: 95°C for 20 s, 65°C for 20 s, 72°C for 50 s.
10. Combine all the first step PCR products and purify a portion using the QIAquick PCR purification Kit (or other column-based purification system).

Time: 2–3 h.
Pause: purified first PCR product can be stored at −20°C for a month as a source for the re-amplification of material in the second PCR.

SECOND PCR AMPLIFICATION

11. Mix the following in a sterile thin-walled tube in a final volume of 25 μl.

Component	Amount, μl	Final concentration
Purified first PCR product	1	
10× polymerase buffer (e.g., Tersus buffer, Evrogen)	2.5	1×
dNTP (2.5 mM each)	1.5	0.15 mM each
Primer Step1 (10 μM)	1	0.4 μM
Reverse primer (10 μM)*	1	0.4 μM
50× polymerase (e.g., Tersus polymerase, Evrogen)	0.5	1×
mQ	17.5	

See **Table 1 for primers used. For primer design options see Sample Barcoding, Unique Molecular Identifiers, and Introducing Diversity at the Ends of the Library in Appendix. In case of simultaneous cDNA synthesis and first PCR amplification of TCR alpha and beta chain libraries, second PCR for TCR alpha and beta chain libraries preparation should be performed in separate reactions. Use an aliquot of purified first PCR product to generate TCR beta library (with beta specific primer) and TCR alpha library (with alpha specific primer).*

Polymerase with high fidelity and processivity should be used for amplification.

12. Carry out amplification using the following program: 95°C for 20 s, 65°C for 20 s, 72°C for 50 s, 9–12 cycles (up to 18–20 cycles if starting from minimal amounts of RNA); final elongation at 72°C for 5 min).
    Purify the PCR products using QIAquick PCR purification Kit (or other column-based purification system) at the same day. This step is important since it removes the residual enzyme activities that can damage the obtained PCR library.

    Time: 2 h.
    Pause: libraries can be stored at −20°C for weeks before adapter ligation.

MIXING THE BARCODED SAMPLES FOR MULTIPLEX SEQUENCING

In order to combine several PCR libraries with pre-introduced sample barcodes (see **Figure 1** and Sample Barcoding in Appendix for possible options), perform the following:

13. Determine the concentration of each library using the QuBit Fluorometer.
14. Combine samples in a sterile microcentrifuge tube proportionally to the desirable amount of sequencing reads per sample. A total amount of PCR products should be approximately 0.5–1 μg (specify the required amount of the PCR product in a sequencing center).

Alternatively, each sample can be ligated to sequencing adapters with different sample barcodes separately. Samples are mixed in desirable proportions before sequencing.

NEXT GENERATION SEQUENCING OPTIONS

Design of the current protocol is optimized for the Illumina paired end 2 × 150 nt (or 2 × 300 nt for IGs) sequencing as the most reliable way to obtain unbiased TCR/IG repertoire. The paired end sequencing is obligatory when double sample barcodes (see and Sample Barcoding in Appendix) and/or unique molecular identifiers (see Unique Molecular Identifiers in Appendix) are used. If no unique molecular identifiers are used, and sample barcoding is used on the 3′-end of the library only (**Figure 1**), then single end sequencing is possible. However, only half of obtained sequencing reads will contain the CDR3 region.

Protocol also suits well the Roche 454 sequencing technology. Frequent length-errors in reading homogenous oligonucleotide stretches on this platform should be kept in mind, and proper error-correction algorithms utilized (10).

In order to use Illumina paired end 2 × 100 nt sequencing for TCRs, the only required modification is that multiplexed J-segment-specific primers should be used instead of the reverse primer in the second PCR amplification step. This minor multiplexing within limited number of PCR cycles does not lead to essential quantitative bias and allows sequence to start closer to the CDR3 region of interest, as described (10, 16). For IG's heavy chain, the universal J-segment-specific primer (**Table 1**) is close to CDR3 already and no modifications are necessary.

Alternative strategy is that sequences for Illumina flow cell and custom sequencing primers can be introduced in the course of amplification (not shown on **Figure 1**). Although potentially beneficial, it requires thorough design in cooperation with sequencing centers.

This protocol is not adopted for Ion Torrent as these sequencing machines have limitations in the maximal length of analyzed sequencing library. Multiplex PCR mix for the V-segment is required for Ion Torrent library preparation, albeit leads to significant quantitative bias during amplification (10).

To provide better cluster differentiation, ask sequencing facility to spike the library with 10–30% of PhiX and/or design primers as described in Introducing Diversity at the Ends of the Library in Appendix.

Size selection on agarose gel after ligation of adapters is strongly recommended since even minor amounts of short non-specific PCR products can significantly reduce target sequences output.

SOFTWARE ANALYSIS OF NGS DATA

Output NGS data on TCR/IG profiling contain numerous errors accumulated during reverse transcription, PCR amplification, and sequencing. For the latter, higher Phred quality score only means lower frequency of sequencing errors. Thus, high sequence quality does not guarantee absence of sequencing errors. Generally, the more we sequence, the more erroneous TCR/IG variants we generate. Without appropriate error-correction, NGS data can generate artificial TCR/IG diversity exceeding the native diversity of complex input library up to several-fold (10).

Several approaches were proposed to correct the PCR and high quality sequencing errors in TCR datasets, suggesting to filter off low frequency TCR variants (8), to filter off the low abundance variants with single mismatch comparing to the major clonotypes (7), or to correct single mismatch errors in germline segments by

Table 1 | Oligonucleotides.

Primer	Application	Sequence*,**
FIRST STRAND cDNA SYNTHESIS		
Switch_oligo	5′ adapter: template switch adapter, universal for all libraries	AAGCAGTGGTATCAACGCAGAGTAC(XXXXX)TCTT(rG)$_5$
SmartNNN	Alternative template switch adapter with unique molecular identifier (see Unique Molecular Identifiers in Appendix), universal for all libraries	AAGCAGUGGTAUCAACGCAGAGUNNNNUNNNNUNNNNUCTT(rG)$_5$
AC1R	Primer for cDNA synthesis, human TCR alpha mRNA	ACACATCAGAATCCTTACTTTG
BC1R	Primer for cDNA synthesis, human TCR beta mRNA	CAGTATCTGGAGTCATTGA
Mus_alfa_synt1	Primer for cDNA synthesis, mouse TCR alpha mRNA	TTTCGGCACATTGATTTG
BC_mus_syn1	Primer for cDNA synthesis, mouse TCR beta mRNA	CAATCTCTGCTTTTGATG
HCA-rt	Primer for cDNA synthesis, human IgA heavy chain mRNA	GTCCGCTTTCGCTCCAGG
HCM-rt	Primer for cDNA synthesis, human IgM heavy chain mRNA	GATGTCAGAGTTGTTCTTG
HCG-rt	Primer for cDNA synthesis, human IgG heavy chain mRNA	GTGTTGCTGGGCTTGTG
FIRST PCR AMPLIFICATION		
Smart20	Step-out primer 1. Anneals on the switch_oligo, universal for all libraries	CACTCTATCCGACAAGCAGTGGTATCAACGCAG
AC2R	Nested primer 1, human TCR alpha library	TACACGGCAGGGTCAGGGT
BC2R	Nested primer 1, human TCR beta library	TGCTTCTGATGGCTCAAACAC
Mus AV2 rev	Nested primer 1, mouse TCR alpha library	GGTGCTGTCCTGAGACCGAG
BC4_mus_Rev	Nested primer 1, mouse TCR beta library	GATGGCTCAAACAAGGAGACC
HCA-n1	Nested primer 1, human IgA heavy chain library	GCGATGACCACGTTCCCATCT
HCM-n1	Nested primer 1, human IgM heavy chain library	GTGATGGAGTCGGGAAGGAAG
HCG-n1	Nested primer 1, human IgG heavy chain library	GAAGTAGTCCTTGACCAGGCA
SECOND PCR AMPLIFICATION		
Step_1	Step-out primer 2, from the Smart20, universal for all libraries	(N)$_{2–4}$(XXXXX)CACTCTATCCGACAAGCAGT
Hum bcj	Nested primer 2, human TCR beta	(N)$_{2–4}$(XXXXX)ACACSTTKTTCAGGTCCTC
Hum acj	Nested primer 2, human TCR alpha	(N)$_{2–4}$(XXXXX)GGGTCAGGGTTCTGGATAT
Mus bcj	Nested primer 2, mouse TCR beta	(N)$_{2–4}$(XXXXX)GGAGTCACATTTCTCAGATCCT
Mus acj	Nested primer 2, mouse TCR alpha	(N)$_{2–4}$(XXXXX)CAGGTTCTGGGTTCTGGATGT
IGHJ-r1	Nested primer 2, human IG heavy chain (universal for IgA, IgG, and IgM)	(N)$_{2–4}$(XXXXX)GAGGAGACGGTGACCRKGGT

XXXXX: optional sample barcode (see **Figure 1, and Sample Barcoding in Appendix for details and Supplementary Material for barcodes). U = dU (deoxyuridine).*
***(N)$_{2–4}$ – optional. Random nucleotides ("N") are introduced at the 5′ end of final library in order to generate diversity for better cluster identification on Illumina sequencer (see Introducing Diversity at the Ends of the Library in Appendix for details).*

mapping to the major clonotypes (10). Low quality sequences can be either filtered off (7, 8) or mapped to the high quality ones in order to rescue quantitative information (10).

There are currently three available software packages for NGS TCR data analysis: IMGT/HighV-QUEST web service[2], Decombinator (28), and our new software, named MiTCR[3] (29). Note that IMGT/HighV-QUEST is limited to only 50,000–150,000 sequences per batch and thus it is hardly suitable for the analysis of deep NGS profiling data. MiTCR is the only software package that considers sequence quality, performs correction of PCR and sequencing errors, and rescues low quality sequencing data. Two basic error-correction modes are currently implemented, aiming either to eliminate maximal number of accumulated errors, or to preserve maximal original TCR diversity, albeit with less efficient error-correction. Moreover, analysis parameters can be tuned by user in a wide range to obtain optimal result for the particular experimental task. Output format is a tab-delimited file or a special *.cls file for the MiTCR-Viewer software (**Figure 2**).

EXPECTED RESULTS

RNA

The quality and quantity of obtained RNA is critical for the library generation. Quality of total RNA is evaluated by two visible bands on electrophoresis (or two highest peaks on Agilent Bioanalyzer) corresponding to 18S and 28S rRNA. The relative amount of two bands should be between 1:2 and 1:1. The expected yield is 1–3 μg of total RNA from one million of PBMC when using Trizol protocol. If starting material is limited (10,000 cells or less) RNA should be completely used in one cDNA synthesis reaction without analyzing by electrophoresis.

NUMBER OF PCR CYCLES

In order to preserve natural TCR/IG diversity of the sample it is important to minimize the number of PCR cycles used for library

[2]http://www.imgt.org/IMGTindex/IMGTHighV-QUEST.html
[3]http://mitcr.milaboratory.com/



FIGURE 2 | MiTCR-viewer outputs for the analyzed TCR beta dataset. **(A)** Table with clonotypes. **(B)** *In silico* spectratyping.

preparation. In our system, maximal number of PCR cycles should be 18 for the first and 12 for the second amplification step if starting from 2 μg of total RNA. A well visible band is observed on electrophoresis after 12 cycles of second PCR amplification (that is at least 50 ng of PCR product per 25 μl reaction). For a minimum amount of starting material (below 10,000 cells) the

maximum number of PCR cycles should be 21 for the first and 18–20 for the second amplification step. If the number of cycles needed to obtain a visible band is higher, this may indicate that low number of molecules has successfully entered amplification, thus leading to uncertain detection of CDR3 clonotypes of the input sample.

SEQUENCING OUTPUT AND ANALYSIS

With the use of the proposed protocol, at least three million of high quality CDR3-containing sequencing reads from a paired end MiSeq run and at least 100 million CDR3-containing sequencing reads from one lane of paired end HiSeq 2,000/2,500 run are expected. The number of different clonotypes depends on the nature and amount of starting material. For example, profiling of 5–10 million human PBMC cells using 1/10 of HiSeq 2000 Illumina lane (at least 10 million CDR3-containg reads) can yield from 0.5 to 2.5 million TCR beta CDR3 clonotypes after appropriate error-correction.

ACKNOWLEDGMENTS

This work was supported by the Molecular and Cell Biology Program RAS, Russian Foundation for Basic Research Grants 12-04-33139, 12-04-00229 (to Dmitriy M. Chudakov), 13-04-00998 (to Olga V. Britanova), 13-04-01124 (to Ilgar Z. Mamedov), Council of the President of the Russian Federation for young scientists СП-2039.2012.4 (to Ivan V. Zvyagin), and European Regional Development Fund CZ.1.05/1.1.00/02.0068.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2013.00456/abstract

REFERENCES

1. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* (2009) **19**:1817–24. doi:10.1101/gr.092924.109
2. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) **114**:4099–107. doi:10.1182/blood-2009-04-217604
3. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, et al. By-passing in vitro screening – next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* (2010) **38**:e193. doi:10.1093/nar/gkq789
4. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* (2010) **28**:965–9. doi:10.1038/nbt.1673
5. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* (2010) **2**:47ra64. doi:10.1126/scitranslmed.3001442
6. Wang C, Sanders CM, Yang Q, Schroeder HW Jr, Wang E, Babrzadeh F, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A* (2010) **107**:1518–23. doi:10.1073/pnas.0913939107
7. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* (2011) **12**:106. doi:10.1186/1471-2164-12-106
8. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
9. Baum PD, Venturi V, Price DA. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* (2012) **42**:2834–9. doi:10.1002/eji.201242999
10. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* (2012) **42**:3073–83. doi:10.1002/eji.201242517
11. Klarenbeek PL, Remmerswaal EB, Ten Berge IJ, Doorenspleet ME, Van Schaik BD, Esveldt RE, et al. Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. *PLoS Pathog* (2012) **8**:e1002889. doi:10.1371/journal.ppat.1002889
12. Linnemann C, Heemskerk B, Kvistborg P, Kluin RJ, Bolotin DA, Chen X, et al. High-throughput identification of antigen-specific TCRs by TCR gene capture. *Nat Med* (2013) **19**:1534–41. doi:10.1038/nm.3359
13. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* (2013) **43**:2507–15. doi:10.1002/eji.201343453
14. Mamedov IZ, Britanova OV, Chkalina AV, Staroverov DB, Amosova AL, Mishin AS, et al. Individual characterization of stably expanded T cell clones in ankylosing spondylitis patients. *Autoimmunity* (2009) **42**:525–36. doi:10.1080/08916930902960362
15. Britanova OV, Staroverov DB, Chkalina AV, Kotlobay AA, Zvezdova ES, Bochkova AG, et al. Single high-dose treatment with glucosaminyl-muramyl dipeptide is ineffective in treating ankylosing spondylitis. *Rheumatol Int* (2011) **31**:1101–3. doi:10.1007/s00296-010-1663-3
16. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, et al. Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol Med* (2011) **3**:201–7. doi:10.1002/emmm.201100129
17. Britanova OV, Bochkova AG, Staroverov DB, Fedorenko DA, Bolotin DA, Mamedov IZ, et al. First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. *Bone Marrow Transplant* (2012) **47**:1479–81. doi:10.1038/bmt.2012.44
18. Luk'ianov SA, Gurskaia NG, Luk'ianov KA, Tarabykin VS, Sverdlov ED. Highly-effective subtractive hybridization of cDNA. *Bioorg Khim* (1994) **20**:701–4.
19. Porter-Jordan K, Rosenberg EI, Keiser JF, Gross JD, Ross AM, Nasim S, et al. Nested polymerase chain reaction assay for the detection of *Cytomegalovirus* overcomes false positives caused by contamination with fragmented DNA. *J Med Virol* (1990) **30**:85–91. doi:10.1002/jmv.1890300202
20. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res* (1999) **27**:1558–60. doi:10.1093/nar/27.6.1558
21. Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL, et al. A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J Immunol* (2002) **168**:3099–104.
22. Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE. Multiplex PCR: optimization and application in diagnostic virology. *Clin Microbiol Rev* (2000) **13**:559–70. doi:10.1128/CMR.13.4.559-570.2000
23. Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal* (2002) **16**:47–51. doi:10.1002/jcla.2058
24. Paillard F, Sterkers G, Vaquero C. Transcriptional and post-transcriptional regulation of TcR, CD4 and CD8 gene expression during activation of normal human T lymphocytes. *EMBO J* (1990) **9**:1867–72.
25. Doskow JR, Wilkinson MF. CD3-gamma, -delta, -epsilon, -zeta, T-cell receptor-alpha and -beta transcripts are independently regulated during thymocyte ontogeny and T-cell activation. *Immunology* (1992) **77**:465–8.
26. Andersen RS, Kvistborg P, Frosig TM, Pedersen NW, Lyngaa R, Bakker AH, et al. Parallel detection of antigen-specific T cell responses by combinatorial encoding of MHC multimers. *Nat Protoc* (2012) **7**:891–902. doi:10.1038/nprot.2012.037
27. Franz B, May KF Jr, Dranoff G, Wucherpfennig K. Ex vivo characterization and isolation of rare memory B cells with antigen tetramers. *Blood* (2011) **118**:348–57. doi:10.1182/blood-2011-03-341917

28. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) **29**:542–50. doi:10.1093/bioinformatics/btt004
29. Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, et al. MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* (2013) **10**:813–4. doi:10.1038/nmeth.2555
30. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* (2012) **9**:72–4. doi:10.1038/nmeth.1778

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 March 2013; accepted: 30 November 2013; published online: 23 December 2013.
*Citation: Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, Lebedev YB and Chudakov DM (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. Front. Immunol. **4**:456. doi: 10.3389/fimmu.2013.00456*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Mamedov, Britanova, Zvyagin, Turchaninova, Bolotin, Putintseva, Lebedev and Chudakov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

SAMPLE BARCODING

When sequencing multiple samples, it is recommended to introduce sample barcodes during the library preparation process. This allows to minimize cross-sample contamination and to treat all samples as the single one when ligating Illumina adapters. It is possible to introduce sample barcodes on different stages (See **Figure 1**). One of the best ways is to use 5′-template switch adapters with built-in sample barcodes, thus labeling each sample at the very first library preparation step. Alternatively/additionally, 5′-end sample barcode can be introduced at the 5′-end of the *Step-out primer 2* (see **Table 1**). We also recommend introduction of sample barcodes within the reverse primers used in the second amplification step (*hum bcj*, *hum acj*, *mus bcj*, *mus acj*, or *IGHJ-r1*, see **Table 1**). Using this approach, each sample is barcoded at both ends of the library. This is crucial when accurate comparison of two or more samples is required, as we observe different levels of swapping ends between molecules in course of standard Illumina library preparation stage and presumably on the solid phase of the sequencer, during bridge amplification. For your convenience, we have generated a list of 5-nucleotide sample barcodes, which differ from each other by at least two nucleotides (see Supplementary Material), thus minimizing the chance of barcode misinterpretation if the single error occurs during sample preparation or sequencing.

UNIQUE MOLECULAR IDENTIFIERS

Unique molecular identifiers can be introduced as random oligonucleotides at the very first amplification (or cDNA synthesis) step of library preparation (30). Each molecule that successfully enters amplification becomes labeled by a unique combination of nucleotides – a molecular identifier. Thus each TCR/IG CDR3 sequence variant in the output NGS dataset is characterized by a number of distinct molecular identifiers indicating the number of such cDNA molecules that have entered the PCR amplification. This approach allows to correct the PCR bias that occurs during amplification and to count mRNA/cDNA molecules of each type directly, which makes the TCR/IG repertoire analysis even more quantitative. Unique molecular identifiers consisting of 12 random nucleotides (which give approximately 17 million unique variants) can be introduced within the 5′-template switch adapter (**Table 1**, SmartNNN). This template switch adapter also contains multiple deoxyuridine nucleotides. After cDNA synthesis, Uracyl DNA glycosylase treatment allows to eliminate SmartNNN, thus preventing possible exchange of unique molecular identifiers during following PCR amplification (30).

INTRODUCING DIVERSITY AT THE ENDS OF THE LIBRARY

The common problem with sequencing PCR products by Illumina is the presence of the same nucleotides in the beginning of most sequencing reads. This can lead to a fail of a sequencing run as Illumina software cannot discriminate adjacent clusters, which produce identical fluorescent signals during the first several sequencing cycles. The common solution used by sequencing centers is spiking the sequencing library by PhiX library containing random DNA fragments. However, in this case, the number of obtained target sequences is decreased by at least 30%. To avoid this problem we introduce two to four random nucleotides ("N") to the 5′ end of the primers used in the second amplification step (see **Table 1**). Preferably, the number of "N" nucleotides flanking the library should be different for the samples mixed on the same Illumina lane, in order to generate additional diversity of starting sequencing steps and to avoid identical nucleotides being present in the same positions, which may alter Illumina sequencing quality. If one sample is sequenced per Illumina lane and no sample barcodes are used, it is recommended to use a mixture of three identical primers, each containing a different number of "N" nucleotides at the 5′ end – e.g., $(N)_2$ Step1/$(N)_3$ Step 1/$(N)_4$ Step1, the same with the reverse primer (see **Table 1**).

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 26 December 2013
doi: 10.3389/fimmu.2013.00485

Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire

Veronika I. Zarnitsyna[1]*, Brian D. Evavold[2], Louis N. Schoettle[3], Joseph N. Blattman[3] and Rustom Antia[1]*

[1] Department of Biology, Emory University, Atlanta, GA, USA
[2] Department of Microbiology and Immunology, Emory University, Atlanta, GA, USA
[3] Center for Infectious Diseases and Vaccinology, School of Life Sciences, Arizona State University, Tempe, AZ, USA

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Mario Castro, Comillas Pontifical University, Spain
Aridaman Pandit, Utrecht University, Netherlands

***Correspondence:**
Veronika I. Zarnitsyna and Rustom Antia, Department of Biology, Emory University, 1510 Clifton Road, Atlanta, GA 30322, USA
e-mail: veronika.i.zarnitsyna@emory.edu; rustom.antia@emory.edu

In order to recognize and combat a diverse array of pathogens the immune system has a large repertoire of T cells having unique T cell receptors (TCRs) with only a few clones specific for any given antigen. We discuss how the number of different possible TCRs encoded in the genome (the potential repertoire) and the number of different TCRs present in an individual (the realized repertoire) can be measured. One puzzle is that the potential repertoire greatly exceeds the realized diversity of naïve T cells within any individual. We show that the existing hypotheses fail to explain why the immune system has the potential to generate far more diversity than is used in an individual, and propose an alternative hypothesis of "evolutionary sloppiness." Another immunological puzzle is why mice and humans have similar repertoires even though humans have over 1000-fold more T cells. We discuss how the idea of the "protecton," the smallest unit of protection, might explain this discrepancy and estimate the size of "protecton" based on available precursor frequencies data. We then consider T cell cross-reactivity – the ability of a T cell clone to respond to more than one epitope. We extend existing calculations to estimate the extent of expected cross-reactivity between the responses to different pathogens. Our results are consistent with two observations: a low probability of observing cross-reactivity between the immune responses to two randomly chosen pathogens; and the ensemble of memory cells being sufficiently diverse to generate cross-reactive responses to new pathogens.

Keywords: $\alpha\beta$ T cell, repertoire, precursor frequency, cross-reactivity, pathogen recognition

1. INTRODUCTION

The clonal selection theory of adaptive immunity requires that the immune system is able to produce a large and diverse repertoire of immune cells (clones), with each cell expressing a receptor with different antigenic specificity (1, 2). Following infection, the few clones that are specific for the antigens expressed by the pathogen proliferate and differentiate into effector cells which control the infection. Subsequently, the maintenance of an increased number of these pathogen-specific cells results in long-lasting immunological memory (3–5). Accurate quantification of changes in the numbers of antigen-specific cells during infection and vaccination, together with advances in molecular and cellular biology, has allowed us to make considerable progress toward understanding the dynamics of the generation of immune responses (3, 6, 7) and the requirements for pathogen control (8, 9). Furthermore, deep sequencing technology has provided a first quantitative snapshot of the diversity of immune cells (10, 11). These technological advances set the stage for understanding the relationship between the diversity of immune cells (the repertoire) and immune protection from an extensive array of pathogens to which we are exposed.

We begin by outlining our current understanding of T cell receptor diversity and discussing problems associated with the quantification of the T cell repertoire. Next, we explore how diverse the immune system needs to be by exploring the relationship between the diversity of the T cell repertoire and its ability to provide protection from pathogens. Finally we consider how the degree of specificity of T cells (often defined by measuring how cross-reactive they are) affects the relationship between the repertoire and host response to a given pathogen. We focus on $\alpha\beta$ T cells and the term "T cell" refers to the CD8 subpopulation of T cells unless we explicitly specify a different subpopulation.

We have intentionally used simple models and calculations because, in the absence of detailed information on the terms and parameters, simpler models frequently generate more robust qualitative results than complex models (12, 13). The focus of the paper is to highlight the limitations arising from uncertainties in current estimates of parameters, and in particular to gain maximum insight from the one key parameter – the precursor frequency of T cells specific for different epitopes – that can be accurately measured. Throughout this paper we emphasize current puzzles and problems and, where possible, suggest new approaches to solving them.

2. MEASURING THE DIVERSITY OF THE T CELL REPERTOIRE

2.1. WHAT IS THE POTENTIAL REPERTOIRE?

T cells develop from progenitor cells in the thymus where the germline T cell receptor (TCR) α and β genes undergo somatic recombination of the V-J and V-D-J gene segments, respectively (14, 15). The antigenic specificity of each T cell is determined

by the amino acid sequence of these rearranged TCR genes, and in particular by the hypervariable complementarity determining region 3 (CDR3) that mostly account for direct contacts with peptides presented on major histocompatibility complex (MHC) proteins, and is encoded by the junction of the V, (D), and J gene segments (16). The diversity of generated TCR genes is therefore due to: (1) selection of one from a number of possible V, D, and J gene segments, (2) semi-random cleavage of recombination hairpin intermediates, and (3) N-nucleotide addition and subtraction at the junction of V, D, and J genes (17, 18). Finally, the pairing of different α and β chains to generate a functional receptor results in additional diversity (19).

How many different T cell receptors can be generated? The first steps toward understanding the magnitude of the diversity of the T cell repertoire came from the pioneering studies that identified the molecular mechanisms involved in the recombination of V, (D), and J gene segments and N-region diversification described above for the generation of the α and β TCR chains (14, 15, 19). It was estimated that these processes together with pairing between different α and β chains could give rise to around 10^{15} possible $\alpha\beta$ TCR (19). The question of the potential number of TCR sequences has recently been revisited and significantly larger estimates for the diversity of the TCR β chain were obtained (20, 21). Murugan et al. (21) used statistical analysis of non-productive TCR β chain to conclude that the CDR3 (variable) region of the TCR β chain alone has a potential diversity of $\sim 10^{14}$ different sequences. They used empirical β chain data to show that N-nucleotide insertions at the V-D and D-J junctions are uncorrelated, their length distributions are nearly identical and their lengths could exceed six nucleotides which was assumed in previous estimates (19). We might expect that a similar analysis would result in upward revision for the potential diversity of the α chain (though the estimates of diversity would increase less than for the β chain because the α chain has only one V-J junction). This will result in a truly enormous potential repertoire of over 10^{20} for the $\alpha\beta$ TCR.

2.2. WHAT IS THE REALIZED REPERTOIRE IN AN INDIVIDUAL?

Only mature T cells that have passed thymic selection (naïve T cells) can be employed in immune responses for protection against pathogens. Thus, in order to understand the balance between diversity and protection, the most important measurement is the "realized" T cell diversity in an individual (i.e., the actual number of different TCR in the mature T cell compartment).

The diversity of the naïve T cell repertoire was initially estimated prior to the advent of deep sequencing technologies by the use of spectrotyping, which involved amplifying mRNA from particular V-J sequence combinations, separating the amplified products on the basis of size, and exhaustive conventional sequencing of a particular length CDR3 product. The diversity of TCR sequences in this sample was then extrapolated to the total T cell population.

2.2.1. TCR diversity and clone size in humans

Arstila et al. (22) used spectrotyping to estimate that there are 10^6 β chains in the blood each pairing on average with at least 25 different α chains, and consequently proposed a lower bound to the estimate of the T cell repertoire in humans of around 2.5×10^7 specificities. Advances in deep sequencing have confirmed that estimation of β chains is in the range of $1 - 4 \times 10^6$ (20, 23, 24).

There is however considerable uncertainty about the extent to which 2.5×10^7 specificities underestimates the diversity of T cells in humans (25, 26). A repertoire of 2.5×10^7 suggests a naïve clone size on average of over 4×10^3 cells ($> 10^{11}/(2.5 \times 10^7)$). This could happen if each clone gets produced multiple times or if once produced a given clone would undergo about 12 rounds of division. The first scenario is unlikely, given the very large estimates of potential diversity (19–21). If the second scenario happens, it must occur in the periphery. Expansion of clones in the thymus would result in a much lower frequency of detectable T cell receptor excision circles (TRECs) in the naïve pool of recent thymic emigrants than is currently observed (27–29). Arstila et al. points out that naïve T cells in the periphery could divide more than 12 times during a human lifespan (26). However, as the total number of naïve T cells remains relatively stable (because division is balanced by death) changes in clone size would have to arise from stochastic drift and this seems unlikely.

2.2.2. TCR diversity and clone size in mice

Interestingly, it was estimated that TCR β chain diversity in mouse spleen is quite similar to the one measured in human blood. The β chain repertoire of $5 - 8 \times 10^5$ specificities with each variable domain of β chain sequence being shared by 30–40 T splenocytes have been reported using spectrotyping technology (30). Pairing with α chain was estimated to add a factor of 2.4 and resulted in total $\alpha\beta$ TCR diversity of 2×10^6. Taking into account that there are 2×10^7 splenocytes it was estimated that the clone size of $\alpha\beta$ TCR is equal to 10 cells (30). The bias in recombination and α-β TCR pairing will likely affect the T cell clone size. A recent study that enumerated the number of naïve T cells specific for different epitopes suggests that there are between 15 and 1500 unique cells in the mouse spleen specific for any given epitope, implying that the number of naïve cells with a given TCR α-β combination is very small, and indeed that most clonotypes have clone size of one (31, 32). This is in contrast with the earlier estimates that suggest an average clone size of 10 cells/clone in the spleen (30). Consequently, it brings the repertoire in the spleen toward the total number of naïve T cells in the spleen, and increases the lower bound of the total $\alpha\beta$ T cell repertoire in the mouse by an order of magnitude. In this case the estimate of 2×10^7 specificities becomes very close to a lower bound estimation for human T cell repertoire.

2.2.3. Limitations in estimates of realized diversity

Current estimates of the realized diversity are lower bounds. The limitations of these studies is the lack of information on the pairing of different TCR α and β chains. Bulk sequencing of a single chain, or even of both TCR α and β chains, is not sufficient to inform us of the potential diversity (33). In principle this problem could be comprehensively addressed by single cell sequencing that would obtain linked α and β chain sequences, but this remains technically and financially infeasible for the large sample sizes required to evaluate naïve repertoires with high diversity (34); the cost of single cell sequencing remains at $1 per cell, making the analysis of T cells from a single mouse more than a $10 million experiment! Oil

emulsion lysis strategies (35) combined with micro-sequencing have increased the capacity of such single-cell studies, but these still are only able to capture <1% of the total naïve T cell repertoire in a single mouse. New techniques or methods need to be developed.

In order to have an accurate and comprehensive quantitative description of diversity, it is important to define what we mean by diversity. We can describe T cell repertoire diversity in terms of summary measures of diversity borrowed from the ecological literature. This includes enumerating the number of distinct clones or computing the Simpson diversity index (36) that takes into account the number of clones and their frequencies. However these summary approaches compress all of the diversity information into a single number. A more comprehensive statistical approach retains the frequency distribution of different clone sizes. In **Figure 1** we show a plot of the frequency distribution of β chain sequences in the mouse naïve T cells using preliminary data. We find a majority of β chain sequences are present at low frequencies and fewer sequences occurring at much higher frequencies. A key problem is that we do not know the α chain sequences pairing with each of these β chains, and this restricts our ability to infer diversity of T cells from these observations.

Several clones in **Figure 1** have very high frequencies and the exact underling mechanisms are not known. The sequences which are more common (generated more frequently) are more likely to be shared between different individuals. It was reported that inbred mice and individuals with the same MHC share some T cells with identical receptors (11, 20, 37, 38). These constitute "public" T cell clones, in contrast with the majority of the T cell clones that are unique to an individual and comprise the "private" part of a repertoire response. In general, the frequency of public TCR clonotypes exceeds what is expected if T cells were chosen at random with equal probability from the total potential repertoire, and perhaps not surprisingly, the clone size of public T cells is higher than that of private T cells in the naïve T cell repertoire [(33); Blattman et al. unpublished results]. This has been suggested to arise due to MHC restriction during thymic selection, biased frequencies of recombination, as well as degeneracy in the genetic code which allows more than one nucleotide sequence to give rise to the same amino acid sequence (33, 39). The factors involved in the evolution and/or selection of public T cell clonotypes and their possible role in the control of infections remain puzzling questions.

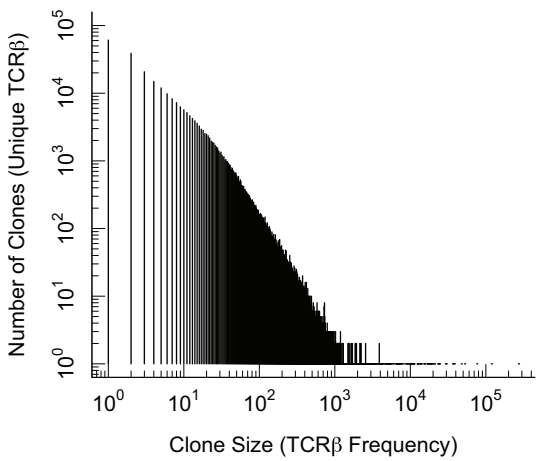


FIGURE 1 | Plot of the frequency distribution in the β chain sequences of naïve CD8 T cells. Naïve (CD44^{lo}) CD8 T cells from C57Bl/6 mice were isolated by magnetic beads and >98% purity confirmed by flow cytometry. Genomic DNA was subjected to TCRβ V-J multiplex DNA sequencing and the distribution of unique in-frame CDR3 sequences is plotted. We note that the term "clone" on the x and y axis labels refers to clones based on β chain sequences alone.

In summary we have estimates of the potential repertoire of upward of 10^{20} TCR. The estimates of the realized repertoire suggest lower bounds of 2.5×10^7 and 2×10^6 in humans and mice. Two puzzles which we will address are: why humans and mice might have similar repertoire sizes (Section 3.2); and why the potential repertoire so greatly exceed the realized repertoire (Section 3.3).

3. UNDERSTANDING DIVERSITY, THE REPERTOIRE AND CROSS-REACTIVITY

In this section we use quantitative calculations to explore the consequences of the observations on the repertoire described in the previous section. We begin by looking at whether the diversity of the repertoire may be explained by the relationship between diversity and protection. We then address questions associated with our current understanding of repertoire diversity and cross-reactivity.

3.1. RELATIONSHIP BETWEEN DIVERSITY AND PROTECTION

Clearly a large repertoire is required to generate a T cell response to a diverse array of pathogens. However, to our knowledge, few empirical studies consider the relationship between the repertoire and protection. To some extent the paucity of experiments on this topic is because of difficulties in quantifying the repertoire (see earlier discussion). Studies on mice, expressing a single fixed transgenic TCR chain (either α or β) that measure the number of different paired endogenously recombined TCR chains, have shown that pairing is not completely random, as these mice express repertoires of reduced diversity and altered V gene usage (40–43). However, even in these mice there is still sufficient diversity to generate effective, albeit reduced, responses to control pathogen infections.

A relatively simple calculation can be made to estimate how diverse the TCR repertoire needs to be in order to provide reliable protection following infection with a pathogen. To provide protection against a pathogen there must be some number of clones present in the repertoire that are specific for that pathogen. Here, we extend the logic outlined in (44, 45). Let R be the T cell repertoire and let p_i be the probability that a randomly chosen TCR binds to i^{th} of the k epitopes derived from a given pathogen ($i = 1{:}k$). Note that p_i is also equal to the precursor frequency of T cells for i^{th} epitope. A pathogen is not detected if all R naïve T cell clones fail to recognize it, and this will happen with probability.

$$P_E = (1-p_1)^R(1-p_2)^R \ldots (1-p_k)^R$$
$$\approx \exp(-p_1 R)\exp(-p_2 R)\ldots \exp(-p_k R) = \exp\left(-R\sum_{i=1}^{k} p_i\right) \quad (1)$$

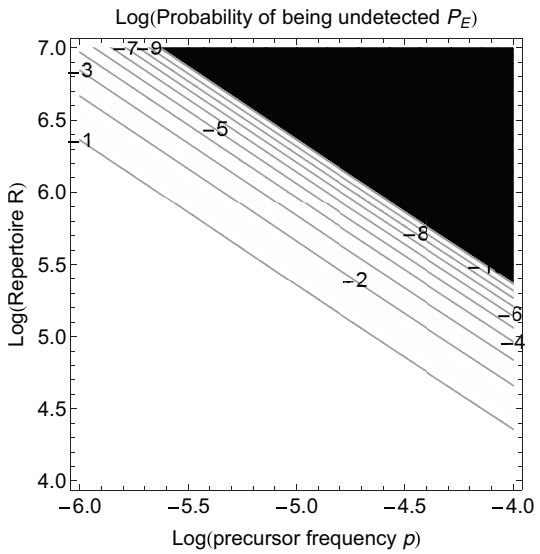


FIGURE 2 | The probability a pathogen is not detected, P_E, as a function of the log of the precursor frequency p and the log of the naïve T cell repertoire R. The numbers on the contour lines in the plot indicate log P_E values. Black color corresponds to the values of P_E below the threshold of 10^{-10}.

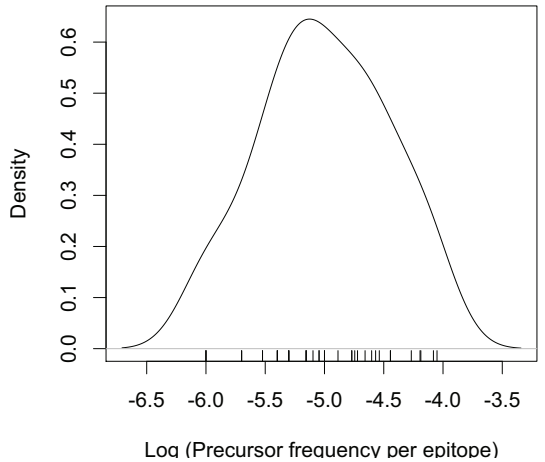


FIGURE 3 | Density distribution plotted from the precursor frequencies of naïve CD8 T cells for different epitopes reported in (31). The tick marks on the top of the x-axis indicate individual epitopes. Note the log scale on the x-axis.

which gives

$$R = \frac{-\ln(P_E)}{\sum_{i=1}^{k} p_i} \tag{2}$$

Equation (2) tells us how big the repertoire must be to detect at level P_E. **Figure 2** shows how the probability that a pathogen is not detected by the immune system depends on the repertoire size and the total precursor frequency $p = \sum p_i$. There is a very rapid decline in the probability of not being detected once the product of p and R becomes sufficiently large. We should note that P_E is often termed as the "probability of escape" but it should not be confused with the usage of the term "escape" that refers to the generation of escape mutants in T cell epitopes after recognition has already occurred following infections such as HIV.

If we know the precursor frequencies for pathogen epitopes and total number of epitopes we can relate the probability of being not detected to the repertoire R. We have much more accurate estimates for precursor frequencies against specific epitopes than for repertoire sizes (31, 46, 47). A recently developed method that combines pMHC tetramer staining with magnetic particle-based cell enrichment allows for the direct measurement of the frequency of naïve cells to different epitopes for both mice and humans (31, 48). **Figure 3** shows the density distribution of naïve T cell precursor frequencies for different CD8 T cell epitopes in mice determined by this cell enrichment method using MHC tetramers complexed with different class I-restricted peptides (31). The total number of responded cells per mouse (naïve precursor frequency multiplied by total CD8 T cell number) varied from 15 in response to the L-338:D^b epitope of LCMV to 1500 in response to an epitope from the murine cytomegalovirus (31). These estimates concur with previous *in vivo* estimates of precursor frequencies. These studies transferred different numbers of naive epitope-specific T cells and measured the proportion of the response arising from expansion of host versus donor cells following virus infection (46, 47). The effect of changing precursor frequencies on the probability of been undetected, P_E, is given by equation (2) and plotted in **Figures 2** and **4**. Note, that precursor frequencies plotted in **Figure 3** are likely biased toward immunodominant epitopes. Immunodominance is a complex issue, and the major factors that affect the magnitude of the T cell response to a particular epitope include: the processing and presentation of peptide on MHC (i.e., the amount of epitope); the number of precursor cells for this epitope; their affinities for the epitope; the extent of their recruitment and competition between the T cells for this and other pathogen epitopes (31, 49–51).

3.2. SCALING AND THE CONCEPT OF A "PROTECTON"

We now consider the scaling of the repertoire with the size of the organism. A few pathogen-specific precursors in a tadpole are likely to be able to mount a faster and more effective response than the same number of cells in an elephant (52). Langman and Cohen proposed the basic functional unit, the "protecton," capable of providing robust protection. They suggested a tadpole (smallest vertebrate) has a single "protecton," and the number of "protectons" scales with the size of an organism. Localized infections are surveyed by a draining lymph node rather than the entire immune system and thus we expect this unit should contain at least one "protecton." Clearly the calculations for P_E (how diverse the immune system needs to be to recognize pathogens) pertains to the "protecton" [see equations (1) and (2)].

Lets estimate the diversity in a "protecton." **Figure 4** shows how the probability of being undetected depends on the size of the repertoire R for different total precursor frequencies. The precursor frequency of T cells specific for a pathogen is, to a first approximation, the sum of the precursor frequencies for that

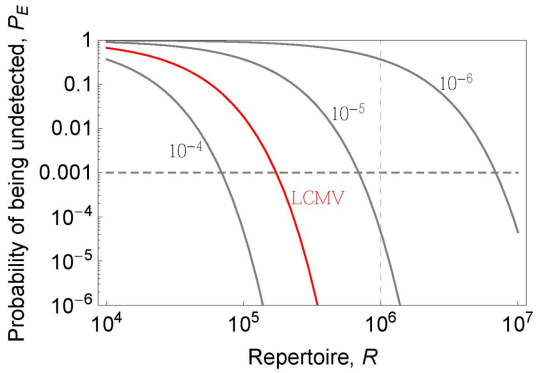


FIGURE 4 | Probability that a pathogen is not recognized P_E (y-axis) is plotted as a function of the repertoire (x-axis) for indicated pathogen-specific precursor frequencies (gray lines). LCMV case ($p = 4 \times 10^{-5}$) is shown in red color.

pathogen's different epitopes presented by MHC proteins. This can be estimated for LCMV by combining reported naïve precursor frequencies for few measured epitopes (31) and measurements of the fraction of the total LCMV specific responses which is directed against these epitopes (53). This gives us a total precursor frequency for LCMV specific T cells equal to $\sim 4 \times 10^{-5}$, and from **Figure 4**, a level of protection $P_E = 10^{-3}$ (i.e., 1 in 1000 "protectons" will fail to recognize LCMV) requires the repertoire in the "protecton" to be about 1.7×10^5. In order to know the level of protection against diverse pathogens we need to know the distribution of precursor frequencies to pathogens. The existing data gives us lower bounds (because only cells specific for a few epitopes are measured) to the precursor frequencies of viruses such as MCMV ($\sim 1.3 \times 10^{-4}$), Influenza ($\sim 4 \times 10^{-5}$), Vaccinia ($\sim 1.1 \times 10^{-4}$), RSV ($\sim 4.5 \times 10^{-5}$), HSV ($\sim 2.9 \times 10^{-5}$), and VSV ($\sim 10^{-5}$) (31). The precursor frequencies for those viruses are comparable or greater than that for LCMV with the exception of VSV and HSV for which only single epitope data were reported in (31). If this trend holds (i.e., the precursor frequency per pathogen is $> 10^{-5}$) it might suggest that having a repertoire of 7×10^{-5} is sufficient to give robust protection at the level $P_E = 10^{-3}$, and thus define the size of the "protecton." For P_E one order of magnitude lower and higher, i.e., $P_E = 10^{-4} - 10^{-2}$ will require a repertoire of $\sim 9 \times 10^5 - 5 \times 10^5$, suggesting that our estimate is quite robust to changes in P_E (see **Figure 4**). We can expect the area of local surveillance (a small lymph node) in mice to have at least this number of different T cells.

How much bigger should the total repertoire size be so that the area corresponding to one "protecton," randomly filled with T cells from the total circulating cells, has a relatively low number of clone repeats? We estimated that if f is a fraction of clone repeats in the "protecton" area with m cells, the total repertoire size R is bounded as $(1-f)(m-1)/(2f) < R < (m-1)/(2f)$. For example, for 5 or 10% of clone repeats in m we will have a multiplication factor for m for the total diversity in the ranges $\sim 9.5 - 10$ or $\sim 4.5 - 5$, respectively. To derive this formula we used two assumptions: first, the clones are equal in size and second, the size of total repertoire multiplied by clone size is much bigger than the size m. These calculations show that the total diversity doesn't need to be much higher than the diversity in a "protecton."

Several theoretical papers previously estimated that the repertoire of B and T cells scales as $ln(cM)$, where c is a constant and M is the mass of an organism (45, 54). It was also estimated that humans should have B cell repertoire 2–5 times larger than mice (45) and similar reasoning could be apply to T cell repertoire. The diversity of the repertoire is linked to clone size and it was estimated that the size of T cell clones should scale as M and, correspondingly, the total number of T cells should scale as $Mln(cM)$ (45). Wiegel and Perelson's estimate shows that the repertoire in a human need not be much higher than that in a mouse even though the number of naïve cells in these organisms differs by over 10^3 fold [mice have $\sim 10^8$ T cells (30, 46) and humans between 10^{11} and 10^{12} T cells (22, 55, 56)].

Another reason for why humans need a more diverse repertoire than mice pertains to the number of pathogens to which they are exposed. As humans live longer than mice, other factors being equal, they will be exposed to more pathogens and require a lower P_E.

3.3. EVOLUTIONARY CONSIDERATIONS: WHY ENCODE SUCH A DIVERSE POTENTIAL REPERTOIRE?

The calculations described in the previous section are consistent with the diversity of the repertoire that is observed in mice and humans (22, 30) (lower bound diversity in the range of 2×10^6 to 2.5×10^7 unique $\alpha\beta$ T cells), and the diversity is sufficient to generate a low probability that a given pathogen is not detected ($P_E < 10^{-4}$). What those estimations don't explain is why the immune system is able to generate a potential diversity of more than 10^{15} T cell specificities (19–21) that is vastly in excess of the realized repertoire?

Let's consider a number of potential explanations for why the potential repertoire needs to be much larger than the realized repertoire. One simplistic explanation takes into account the observation that most of the generated progenitor T cells are deleted during positive and negative selection in the thymus. If a fraction f of the T cells generated in the thymus gets selected (i.e., pass positive or negative selection) then the potential repertoire should be $(1/f)$ times the peripheral repertoire. Since only 3–5% of T cells pass thymic selection (57, 58), the potential repertoire need only be at most 33-fold higher than the realized repertoire, thus ruling out this explanation.

A second potential reason is the need to successfully recognize peptides in the context of the hundreds of MHC alleles in the population. The reported extent of thymic selection (see previous paragraph) allows us to reject this hypothesis – different cells may be selected in different MHC backgrounds but in all cases 3–5% of T cells pass thymic selection.

A third potential reason is the need to prevent pathogen escape mutations – mutations in an epitope that prevent it from being recognized by the immune system. To a first approximation having more than one epitope is the key factor that prevents escape – if the pathogen has k epitopes the probability of escaping all epitopes declines as μ^k, where μ is the probability of mutation leading to the loss of one epitope. The number of epitopes to which a response is generated involves many factors such as immunodominance,

MHC diversity, and T cell diversity. The relationship between these quantities is not understood and we do not know the contribution of T cell diversity to immunodominance due to problems in estimating TCR diversity described above. However from **Figure 4** we note that the repertoire is sufficiently large to enable robust detection of subdominant epitopes in a biologically reasonable range of precursor frequencies [**Figure 2**; (31)].

A fourth potential reason considers the temporal aspect and changes in the repertoire over the lifespan of an individual. Thymic emigration results in a constantly changing repertoire over time. The total number of different T cells present in the individual over its lifespan could be much greater than its repertoire at any given time. In humans, for example, if we assume that thymic emigration is of the order of $10^7 - 10^8$ cells per day (59, 60) then the realized repertoire over a lifespan might be as much as 10^{12} specificities which is much closer to the potential repertoire. There are two problems with this approach. First, it does not explain why mice have about the same potential repertoire as humans since a similar calculation for mice would result in a realized repertoire over the lifespan several orders of magnitude lower than humans. This is because both mice thymic output of the order of 10^6 cells per day (61–63) and lifespan are smaller than for humans. Second, protection is related to the repertoire at a given time point. Changing the particular clones in the repertoire over time does not help unless the relevant clones are present at the time of infection or generated during the infection and consequently able to help with clearance of the pathogen. The continual influx of cells of new specificities is thus unlikely to be of significance for acute infections which are relatively brief, but has been suggested to contribute to the maintenance of the response during persistent infections (64–66). In the case of persistent infections, however, an occasional new pathogen-specific clone is unlikely to clear the infection if the much larger number of clones at the onset were not able to do so – and the new clone is likely to be exhausted rapidly. Finally, temporal aspects could change the total repertoire if we consider the sum of both naïve and memory compartments. As naïve cells convert to memory cells each time we confront an infection, the replenishment of naïve compartment with the cells of new specificities would increase the total repertoire (naïve plus memory compartments). However, even if the memory compartment is as diverse as the naïve, the total diversity would increase at most by a factor of 2 in comparison to the naïve compartment alone. Taken together we don't expect temporal aspects to account for the differences in the sizes of the potential and realized repertoires.

We now describe a new evolutionary explanation that we call "evolutionary sloppiness." The process of generation of diversity by recombination and N nucleotide addition and deletion are sloppy processes. To reduce the amount of diversity that can be generated might require putting additional costly constraints on these processes. This would explain why organisms are able to generate far more TCR diversity (in excess of 10^{15} TCR) then is needed. Finally we note that not all aspects of biology result in perfectly optimized solutions (67).

Additionally, it has been suggested that the thymus is an energy- and resource-expensive organ (68) but these energetic costs have yet to be quantified. Energetic costs to cell production and thymic selection would favor expansion of clones after thymic selection (i.e., to have an amplifier). This amplification could take place in the thymus or periphery and would scale with the size of the organism. This would result in clone sizes in men being ~1000-fold higher than in mice, which is unlikely (see discussion on TRECs in Section 2.2.1). Accurate estimates for clone sizes in humans and mice should allow us to resolve this question.

3.4. T CELL CROSS-REACTIVITY

Cross-reactivity is related to the observation that a given T cell can respond to more than one epitope, including epitopes that show strong sequence homology or completely unrelated (69–76). As might be expected the frequency of the former is higher than that of the latter. Flexible TCR-pMHC binding sites were suggested as a possible structural explanation for known high degree of $\alpha\beta$ TCRs cross-reactivity to different pMHCs (77–80). Cross-reactivity can also arise in T cell clones with incomplete allelic exclusion at the α chain loci resulting in one β chain pairing with two different α chains. An upper bound on the frequency of such clones was estimated to be 30% (81–83).

The pioneering experiments of Selin and Welsh (69) found that the CD8 T cell responses of mice to pathogens such as the Pichinde virus (PV), Vaccinia virus (VV), and Lymphocytic choriomeningitis virus (LCMV) showed high levels of cross-reactivity. They found that prior vaccination with one of these viruses expanded a specific CD8 T cell subset that could be boosted during stimulation by the other viruses and showed an unexpectedly complex relationship between the responses to different viruses with asymmetry depending on the order of viral exposure (infection A followed by B stimulated different cross-reactivity than B followed by A) (69, 71, 84). In a very recent paper (80) the cross-reactivity studies were extended to analysis of the CD4 T cell repertoire against pathogens to which individuals had never been exposed. Surprisingly, they found that a large fraction of the CD4 T cells specific for these pathogens exhibited a memory phenotype and suggested that they had been generated by cross-reactive responses to other previously encountered pathogens including heterologous infections or environmental antigens.

The extent of cross-reactivity between the immune responses to different pathogens is of practical importance. Murine studies have not only demonstrated the presence of T cells that could cross-react between different pathogens such as PV, VV, and LCMV, but also showed that this cross-reactivity affected pathogenesis during subsequent infection (84–86). If these occurrences of cross-reactive responses are rare, then the examples above are simply interesting curiosities. If, on the other hand, cross-reactivity is common then we would need to move from our current paradigm, which looks at each infection independent of other infections, to a more complex view that incorporates the terms for the interactions between the immune responses to different pathogens. Thus a key step is to quantify the extent of cross-reactivity in the immune responses to different pathogens.

How can we predict the level of cross-reactivity between two pathogens? The current approach is based on the observation that number of possible peptide-MHC complexes is much larger than the total number of T cells, suggesting that a given T cell must be able to recognize many different peptide-MHC (i.e. have a high level of cross-reactivity) (87, 88). However it is not clear how these

parameters can be measured. We propose an alternative calculation that allows estimation of the extent of cross-reactivity from the precursor frequency of T cells for pathogens – a parameter that can be reliably determined. Using slightly modified notation from (87) we first define four parameters:

R Repertoire (the number of clonotypically different naïve T cells in the repertoire)
r The number of different T cell clonotypes that will respond to the same peptide
N The total number of potentially immunogenic foreign peptides in the environment
n The number of different peptides to which a single T cell clonotype will respond

These four parameters are linked by the conservation equation (87):

$$rN = nR \tag{3}$$

Lets suggest that a given pathogen has k epitopes to which T cells can mount a response. For a given T cell, the probability to recognize at least one epitope from a given pathogen could be written as:

$$1 - \left[1 - \frac{n}{N}\right]^k \tag{4}$$

The probability that the same T cell will recognize at least one epitope on each of two pathogens with k epitopes (i.e., will be cross-reactive) will be a square of expression equation (4) and the probability to find at least one cross-reactive clone is equal to:

$$\left[1 - \left[1 - \frac{n}{N}\right]^k\right]^2 \times R \tag{5}$$

Note, that in derived equation (5) we don't know the parameters k, n, and N which makes it very difficult to apply directly. Interestingly, for one epitope ($k = 1$) and with application of cross-reactivity equation (3) the equation (5) simplifies to:

$$\left[1 - \left[1 - \frac{n}{N}\right]\right]^2 \times R = \left[1 - \left[1 - \frac{r}{R}\right]\right]^2 \times R = \left[\frac{r}{R}\right]^2 \times R \tag{6}$$

Under the assumption that all naïve T cells able to respond to a given epitope are clonotypically different, which is supported by recent data (31, 80), we can think of r/R as a precursor frequency for a given epitope. The problem of estimating cross-reactivity in this case will be similar to the problem of estimating the probability of randomly choosing a two-colored ball from an urn when the frequencies of each of two colors are known. Interestingly, the measured naïve precursor frequencies for different immunogenic epitopes are similar for mice and humans and range from 1 to 100 cells per million cells (31). According to equation (3), this similarity immediately implies that cross-reactivity of each T cell receptor, n, is in the same range for mice and humans.

For the case when $k > 1$, we could not directly use the formula (5), due to unknown parameters, but can use a simple probabilistic calculation based on sampling multiple colored balls. We can write the frequency of cross-reactive cells between two randomly chosen pathogens (A and B) in terms of the precursor frequencies of T cells to these two pathogens, p_A and p_B, and the total number of cells T:

$$\text{Expected number of cross-reactive cells} = p_A p_B \times T \tag{7}$$

and if we have clones of same size equal to T/R,

$$\text{Expected number of cross-reactive clones} = p_A p_B \times R \tag{8}$$

where R equals the repertoire (the number of different T cell clones in each individual). When the average number of cross-reactive clones is less than one, equation (8) gives us the probability of observing a cross-reactive response between two pathogens in a single individual. We can use this framework together with the data on precursor frequencies (31) described in **Figure 2** to get an estimate of the extent of cross-reactivity between the responses to unrelated pathogens. As described earlier we have an approximate precursor frequency per epitope of 10^{-5}, and a precursor frequency per pathogen, with LCMV as an example, of about 4×10^{-5}. If we assume there are about 2×10^7 naïve CD8 T cells per mouse (89) then the number of cross-reactive cells between two unrelated pathogens will be ~0.032, which suggests cross-reactivity is very rare (a single cross-reactive clone will be found <4% of the time). In order to observe cross-reactivity between two random pathogens in a mouse we would need to have a precursor frequency per pathogen of at least 2.2×10^{-4}, assuming that precursor frequencies are similar for both pathogens. There are quite a few reported examples of cross-reactive T cell responses to different pathogens. In addition to the experiments of Selin and Welsh (69), cross-reactivity has been reported between influenza virus and hepatitis C virus (72), EBV (73) or HIV (74), LCMV and vaccinia virus (75), and coronavirus and human papillomavirus (76). It remains to be seen if the observed cases of cross-reactivity arise from a reporting bias (failure to observe cross-reactivity between two pathogens is unlikely to be reported) or because some of the assumptions of our model are incorrect and need to be modified. For example, we assume all T cell clones have the same level of cross-reactivity – and introducing heterogeneity may dramatically increase the chances to observe cross-reactivity.

Even if cross-reactive T cell responses to two specific pathogens are rare, the accumulation of many successive infections could result in fairly frequent cross-reactivity between a new pathogen and sum of all the pathogens the individual has previously encountered. This is what was observed when T cell precursor frequencies were measured for novel pathogens in blood from adults (80). The precursor frequency for cells recognizing a self-antigen or an unexposed viral epitope was the same as earlier estimated in mice and humans (31) and ranged from one to ten cells per million of CD4 T cells. The surprise was that over half of the precursor cells specific for novel pathogens such as HIV (to which an individual had never been exposed) were of the memory phenotype (80), suggesting that they may have arisen as a consequence of exposure to a different previously encountered pathogen(s). Alternatively, these memory cells could be pseudo-memory cells acquired via the process known as "homeostatic proliferation" and driven by

interaction with low-affinity self pMHCs that previously induced positive selection (90–94).

The Su et al. paper (80) raised an interesting question: why do memory cells invariably contribute about 50–80% of the precursors to pathogen the individual has never encountered? One possibility is that the memory repertoire is sufficiently large to be "complete." In this case if we draw the same amount of cells from either naïve or memory compartments (or mixture from both) we will have the same precursor frequency for a pathogen. Then the relative contribution of naïve and memory cells to precursors is equal to their relative frequencies, and is scaled by the stimulation threshold which is known to be lower for memory cells.

We note that our equation (7) allows an estimation of cross-reactivity for unrelated pathogens or peptides and, based on reported precursor frequencies for different epitopes, we expect cross-reactivity to be rare. Several studies allowed to estimate the rate of cross-reactivity for closely related peptides. Su et al. (80) identified potential pathogens responsible for generating T cells cross-reactive to HIV in HIV-negative individuals as follows: they generated clones from the HIV precursors and identified two epitopes to which these clones were specific. Then using a BLAST search of pathogen sequences they identified 24 sequences similar to the two HIV epitopes. About 21% of the HIV clones responded to two of the BLAST sequences corresponding to environmental pathogens. This number is comparable to result obtained in the earlier study, which showed that although the majority of 171 generated variant peptides of strongly immunodominant HLA-A2-restricted HIV gag epitope were able to bind HLA-A2, only one third were recognized by specific T cells (95). These two studies may give the rate of cross-reactivity for closely related peptides (21–33%) and could be particularly important in the context of a variable virus with an increased rate of mutations within epitopes (96).

Cross-reactive responses may be of clinical importance in the generation of pathology and autoimmunity. Several studies pointed that cross-reactivity may be the cause of increased immunopathology during successive unrelated viral infections (84–86) or as a result of application of T cell based therapy (97, 98). Expansion of cross-reactive T cell clones due to previous infections may underlie autoimmune diseases (99–101). Sometimes a pathogen epitope stimulates T cells in the context of a different MHC from the self-epitopes that react with these T cells, for example, Epstein-Barr virus EBN13-HLA-B8-specific cytotoxic T cells were shown to cross-react with a variety of self peptides presented by HLA-B35 (102). Together these observations point out that cross-reactive T cell responses might operate on different levels and much remains to be done to understand the extent of cross-reactivity and how it may differ in CD8, CD4, and regulatory populations of T cells.

4. SUMMARY

We have reviewed current estimates of the T cell repertoire and identified their key limitations. Further progress will require the development of methods to determine the pairing of TCR α and β chains and thus more accurate quantification of the T cell diversity. Current estimates raise the puzzling question of why the potential repertoire is many orders of magnitude greater than the realized repertoire. We suggest that existing hypotheses are not able to explain this puzzle and have proposed an alternative hypothesis of "evolutionary sloppiness."

One of the interesting observation that became obvious from our estimations is that precursor frequency per pathogen inherently links the TCR diversity and cross-reactivity which allows to predict the level of cross-reactivity between two random pathogens or unrelated peptides. Our estimates suggest that although cross-reactivity is a rare event for immunologically naïve individuals, probability to see the cross-reactive memory T cells becomes very high with an increase in successive infections.

Finally we note that we need to move from our current paradigm, which looks at each infection independent of other infections, to a more complex view that incorporates the terms for the interactions between the immune responses to different pathogens.

ACKNOWLEDGMENTS

Veronika I. Zarnitsyna thanks the organizers of the Multi-Scale Physics of Lymphocyte Development Summer Seminar 2012 that sparked her interest in the problem of T cell cross-reactivity. Brian D. Evavold was supported by NIH R01 NS071518 and R01 AI096879, Joseph N. Blattman was supported by a Virginia G. Piper Personalized Medicine Bridging Award, and Veronika I. Zarnitsyna and Rustom Antia were supported by an NIH MIDAS grant U01 GM070749. We also thank the Reviewers for their comments and suggestions.

REFERENCES

1. Burnet F. A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA Cancer J Clin* (1957) **26**:119–21. doi:10.3322/canjclin.26.2.119
2. Talmage DW. Allergy and immunology. *Annu Rev Med* (1957) **8**:239–56. doi:10.1146/annurev.me.08.020157.001323
3. Murali-Krishna K, Altman J, Suresh M, Sourdive D, Zajac A, Miller J, et al. Counting antigen-specific CD8+ T cells: a reevaluation of bystander activation during viral infection. *Immunity* (1998) **8**:177–87. doi:10.1016/S1074-7613(00)80470-7
4. Jamieson B, Ahmed R. T cell memory: long term persistence of virus specific cytotoxic cells. *J Exp Med* (1989) **169**:1993–2005. doi:10.1084/jem.169.6.1993
5. Miller JF, Sprent J. Cell-to-cell interaction in the immune response. VI. Contribution of thymus-derived cells and antibody-forming cell precursors to immunological memory. *J Exp Med* (1971) **134**(1):66–82. doi:10.1084/jem.134.1.66
6. Perelson AS. Modeling viral and immune system dynamics. *Nat Rev Immunol* (2002) **2**(1):28–36. doi:10.1038/nri700
7. Antia R, Ganusov VV, Ahmed R. The role of models in understanding CD8+ T-cell memory. *Nat Rev Immunol* (2005) **5**(2):101–11. doi:10.1038/nri1550
8. Regoes RR, Barber DL, Ahmed R, Antia R. Estimation of the rate of killing by cytotoxic T lymphocytes in vivo. *Proc Natl Acad Sci USA* (2007) **104**(5):1599–603. doi:10.1073/pnas.0508830104
9. Yates AJ, Van Baalen M, Antia R. Virus replication strategies and the critical CTL numbers required for the control of infection. *PLoS Comp Biol* (2011) **7**(11):e1002274. doi:10.1371/journal.pcbi.1002274
10. Sherwood AM, Desmarais C, Livingston RJ, Andriesen J, Haussler M, Carlson CS, et al. Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after $\alpha\beta$ and $\gamma\delta$ T cell commitment. *Sci Transl Med* (2011) **3**(90):90ra61.
11. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* (2011) **186**(7):4285–94. doi:10.4049/jimmunol.1003898

12. Hilborn R, Mangel M. The ecological detective: confronting models with data. In: Hilborn R, Mangel M, editors. *Monographs in Population Biology*. Princeton, NJ: Princeton University Press (1997). 28 p.
13. May RM. Uses and abuses of mathematics in biology. *Science* (2004) **303**(5659):790–3. doi:10.1126/science.1094442
14. Kurosawa Y, von Boehmer H, Haas W, Sakano H, Trauneker A, Tonegawa S. Identification of d segments of immunoglobulin heavy-chain genes and their rearrangement in T lymphocytes. *Nature* (1981) **290**(5807):565–70. doi:10.1038/290565a0
15. Hayday AC, Saito H, Gillies SD, Kranz DM, Tanigawa G, Eisen HN, et al. Structure, organization, and somatic rearrangement of T cell gamma genes. *Cell* (1985) **40**(2):259–69. doi:10.1016/0092-8674(85)90140-0
16. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* (1996) **384**(6605):134–41. doi:10.1038/384134a0
17. Oltz EM. Regulation of antigen receptor gene assembly in lymphocytes. *Immunol Res* (2001) **23**(2–3):121–33. doi:10.1385/IR:23:2-3:121
18. Jolly CJ, Cook AJL, Manis JP. Fixing DNA breaks during class switch recombination. *J Exp Med* (2008) **205**(3):509–13. doi:10.1084/jem.20080356
19. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* (1988) **334**(6181):395–402. doi:10.1038/334395a0
20. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* (2010) **2**(47):47ra64. doi:10.1126/scitranslmed.3001442
21. Murugan A, Mora T, Walczak AM, Callan CG Jr. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* (2012) **109**(40):16161–6. doi:10.1073/pnas.1212755109
22. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. *Science* (1999) **286**(5441):958–61. doi:10.1126/science.286.5441.958
23. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) **114**(19):4099–107. doi:10.1182/blood-2009-04-217604
24. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**(5):790–7. doi:10.1101/gr.115428.110
25. Kesmir C, Borghans JA, de Boer RJ. Diversity of human $\alpha\beta$ T cell receptors. *Science* (2000) **288**:1135. doi:10.1126/science.288.5469.1135a
26. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. Diversity of human alpha beta T cell receptors. *Science* (2000) **288**(5469):1135. doi:10.1126/science.288.5469.1135a
27. Zhang L, Lewin SR, Markowitz M, Lin HH, Skulsky E, Karanicolas R, et al. Measuring recent thymic emigrants in blood of normal and HIV-1-infected individuals before and after effective therapy. *J Exp Med* (1999) **190**(5):725–32. doi:10.1084/jem.190.5.725
28. Bains I, Antia R, Callard R, Yates AJ. Quantifying the development of the peripheral naive CD4+ T-cell pool in humans. *Blood* (2009) **113**(22):5480–7. doi:10.1182/blood-2008-10-184184
29. McFarland RD, Douek DC, Koup RA, Picker LJ. Identification of a human recent thymic emigrant phenotype. *Proc Natl Acad Sci U S A* (2000) **97**(8):4215–20. doi:10.1073/pnas.070061597
30. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. *J Immunol* (2000) **164**(11):5782–7.
31. Jenkins MK, Moon JJ. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. *J Immunol* (2012) **188**(9):4135–40. doi:10.4049/jimmunol.1102661
32. Tubo NJ, Pagán AJ, Taylor JJ, Nelson RW, Linehan JL, Ertelt JM, et al. Single naive CD4(+) T cells from a diverse repertoire produce different effector cell types during infection. *Cell* (2013) **153**(4):785–96. doi:10.1016/j.cell.2013.04.007
33. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T cell responses? *Nat Rev Immunol* (2008) **8**(3):231–8. doi:10.1038/nri2260
34. Dash P, McClaren JL, Oguin TH III, Rothwell W, Todd B, Morris MY, et al. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J Clin Invest* (2011) **121**(1):288–95. doi:10.1172/JCI44752
35. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* (2013) **43**(9):2507–15. doi:10.1002/eji.201343453
36. Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* (2007) **321**(1–2):182–95. doi:10.1016/j.jim.2007.01.019
37. Bousso P, Casrouge A, Altman JD, Haury M, Kanellopoulos J, Abastado JP, et al. Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* (1998) **9**(2):169–78. doi:10.1016/S1074-7613(00)80599-3
38. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) **107**(45):19414–9. doi:10.1073/pnas.1010586107
39. Livak F, Burtrum DB, Rowen L, Schatz DG, Petrie HT. Genetic modulation of T cell receptor gene segment usage during somatic recombination. *J Exp Med* (2000) **192**(8):1191–6. doi:10.1084/jem.192.8.1191
40. Brändle D, Bürki K, Wallace VA, Rohrer UH, Mak TW, Malissen B, et al. Involvement of both T cell receptor V alpha and V beta variable region domains and alpha chain junctional region in viral antigen recognition. *Eur J Immunol* (1991) **21**(9):2195–202. doi:10.1002/eji.1830210930
41. Perkins DL, Wang YS, Fruman D, Seidman JG, Rimm IJ. Immunodominance is altered in T cell receptor (beta-chain) transgenic mice without the generation of a hole in the repertoire. *J Immunol* (1991) **146**(9):2960–4.
42. Turner SJ, Cose SC, Carbone FR. TCR alpha-chain usage can determine antigen-selected TCR beta-chain repertoire diversity. *J Immunol* (1996) **157**(11):4979–85.
43. Burns RP Jr, Natarajan K, LoCascio NJ, O'Brien DP, Kobori JA, Shastri N, et al. Molecular analysis of skewed Tcra-V gene use in T-cell receptor beta-chain transgenic mice. *Immunogenetics* (1998) **47**(2):107–14. doi:10.1007/s002510050335
44. De Boer RJ, Perelson AS. How diverse should the immune system be? *Proc R Soc Lond B Biol Sci* (1993) **252**(1335):171–5. doi:10.1098/rspb.1993.0062
45. Wiegel FW, Perelson AS. Some scaling principles for the immune system. *Immunol Cell Biol* (2004) **82**(2):127–31. doi:10.1046/j.0818-9641.2004.01229.x
46. Blattman JN, Antia R, Sourdive DJ, Wang X, Kaech SM, Murali-Krishna K, et al. Estimating the precursor frequency of naive antigen-specific CD8 T cells. *J Exp Med* (2002) **195**(5):657–64. doi:10.1084/jem.20001021
47. Whitmire JK, Benning N, Whitton JL. Precursor frequency, nonlinear proliferation, and functional maturation of virus-specific CD4+ T cells. *J Immunol* (2006) **176**(5):3028–36.
48. Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, et al. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* (2007) **27**(2):203–13. doi:10.1016/j.immuni.2007.07.007
49. Thomas PG, Handel A, Doherty PC, La Gruta NL. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naive and immune T-cell populations. *Proc Natl Acad Sci U S A* (2013) **110**(5):1839–44. doi:10.1073/pnas.1222149110
50. La Gruta NL, Rothwell WT, Cukalac T, Swan NG, Valkenburg SA, Kedzierska K, et al. Primary CTL response magnitude in mice is determined by the extent of naive T cell recruitment and subsequent clonal expansion. *J Clin Invest* (2010) **120**(6):1885–94. doi:10.1172/JCI41538
51. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol* (1999) **17**:51–88. doi:10.1146/annurev.immunol.17.1.51
52. Langman R, Cohen M. The E-T (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protection. *Mol Immunol* (1987) **24**:675–97. doi:10.1016/0161-5890(87)90050-2
53. Masopust D, Murali-Krishna K, Ahmed R. Quantitating the magnitude of the lymphocytic choriomeningitis virus-specific CD8 T-cell response: it is even bigger than we thought. *J Virol* (2007) **81**(4):2002–11. doi:10.1128/JVI.01459-06
54. Perelson AS, Wiegel FW. Scaling aspects of lymphocyte trafficking. *J Theor Biol* (2009) **257**(1):9–16. doi:10.1016/j.jtbi.2008.11.007
55. Westermann J, Pabst R. Lymphocyte subsets in the blood: a diagnostic window on the lymphoid system? *Immunol Today* (1990) **11**(11):406–10. doi:10.1016/0167-5699(90)90160-B

56. Clark DR, de Boer RJ, Wolthers KC, Miedema F. T cell dynamics in HIV-1 infection. *Adv Immunol* (1999) **73**:301–27. doi:10.1016/S0065-2776(08)60789-0
57. Shortman K, Vremec D, Egerton M. The kinetics of T cell antigen receptor expression by subgroups of CD4+8+ thymocytes: delineation of CD4+8+3(2+) thymocytes as post-selection intermediates leading to mature T cells. *J Exp Med* (1991) **173**(2):323–32. doi:10.1084/jem.173.2.323
58. Huesmann M, Scott B, Kisielow P, von Boehmer H. Kinetics and efficacy of positive selection in the thymus of normal and T cell receptor transgenic mice. *Cell* (1991) **66**(3):533–40. doi:10.1016/0092-8674(81)90016-7
59. Vrisekoop N, den Braber I, de Boer AB, Ruiter AFC, Ackermans MT, van der Crabben SN, et al. Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool. *Proc Natl Acad Sci U S A* (2008) **105**(16):6115–20. doi:10.1073/pnas.0709713105
60. Hazenberg MD, Borghans JA, de Boer RJ, Miedema F. Thymic output: a bad TREC record. *Nat Immunol* (2003) **4**(2):97–9. doi:10.1038/ni0203-97
61. Tough D, Sprent J. Turnover of naive- and memory-phenotype T cells. *J Exp Med* (1994) **179**:1127–35. doi:10.1084/jem.179.4.1127
62. Scollay RG, Butcher EC, Weissman IL. Thymus cell migration quantitative aspects of cellular traffic from the thymus to the periphery in mice. *Eur J Immunol* (1980) **10**(3):210–8. doi:10.1002/eji.1830100310
63. Gabor MJ, Scollay R, Godfrey DI. Thymic T cell export is not influenced by the peripheral T cell pool. *Eur J Immunol* (1997) **27**(11):2986–93. doi:10.1002/eji.1830271135
64. Vezys V, Masopust D, Kemball CC, Barber DL, O'Mara LA, Larsen CP, et al. Continuous recruitment of naive T cells contributes to heterogeneity of antiviral CD8 T cells during persistent infection. *J Exp Med* (2006) **203**(10):2263–9. doi:10.1084/jem.20060995
65. Yang Y, Al-Mozaini M, Buzon MJ, Beamon J, Ferrando-Martinez S, Ruiz-Mateos E, et al. CD4 T-cell regeneration in HIV-1 elite controllers. *AIDS* (2012) **26**(6):701–6. doi:10.1097/QAD.0b013e3283519b22
66. Hatzakis A, Touloumi G, Karanicolas R, Karafoulidou A, Mandalaki T, Anastassopoulou C, et al. Effect of recent thymic emigrants on progression of HIV-1 disease. *Lancet* (2000) **355**(9204):599–604. doi:10.1016/S0140-6736(99)10311-8
67. Gould S, Lewontin R. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* (1979) **205**:581–98. doi:10.1098/rspb.1979.0086
68. George A, Ritter M. Thymic involution with ageing: obsolescence or good housekeeping? *Immunol Today* (1996) **17**:267–72. doi:10.1016/0167-5699(96)80543-3
69. Selin LK, Nahill SR, Welsh RM. Cross-reactivities in memory cytotoxic T lymphocyte recognition of heterologous viruses. *J Exp Med* (1994) **179**(6):1933–43. doi:10.1084/jem.179.6.1933
70. Selin LK, Brehm MA, Naumov YN, Cornberg M, Kim S-K, Clute SC, et al. Memory of mice and men: CD8+ T-cell cross-reactivity and heterologous immunity. *Immunol Rev* (2006) **211**:164–81. doi:10.1111/j.0105-2896.2006.00394.x
71. Welsh RM, Selin LK. No one is naive: the significance of heterologous T-cell immunity. *Nat Rev Immunol* (2002) **2**(6):417–26.
72. Wedemeyer H, Mizukoshi E, Davis AR, Bennink JR, Rehermann B. Cross-reactivity between hepatitis c virus and influenza a virus determinant-specific cytotoxic T cells. *J Virol* (2001) **75**(23):11392–400. doi:10.1128/JVI.75.23.11392-11400.2001
73. Clute SC, Watkin LB, Cornberg M, Naumov YN, Sullivan JL, Luzuriaga K, et al. Cross-reactive influenza virus-specific CD8+ T cells contribute to lymphoproliferation in Epstein-Barr virus-associated infectious mononucleosis. *J Clin Invest* (2005) **115**(12):3602–12. doi:10.1172/JCI25078
74. Acierno PM, Newton DA, Brown EA, Maes LA, Baatz JE, Gattoni-Celli S. Cross-reactivity between HLA-A2-restricted FLU-M1:58-66 and HIV p17 GAG:77-85 epitopes in HIV-infected and uninfected individuals. *J Transl Med* (2003) **1**(1):3. doi:10.1186/1479-5876-1-3
75. Kim S-K, Cornberg M, Wang XZ, Chen HD, Selin LK, Welsh RM. Private specificities of CD8 T cell responses control patterns of heterologous immunity. *J Exp Med* (2005) **201**(4):523–33. doi:10.1084/jem.20041337
76. Nilges K, Höhn H, Pilch H, Neukirch C, Freitag K, Talbot PJ, et al. Human papillomavirus type 16 E7 peptide-directed CD8+ T cells from patients with cervical cancer are cross-reactive with the coronavirus NS2 protein. *J Virol* (2003) **77**(9):5464–74. doi:10.1128/JVI.77.9.5464-5474.2003
77. Reinherz EL, Tan K, Tang L, Kern P, Liu J, Xiong Y, et al. The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science* (1999) **286**(5446):1913–21. doi:10.1126/science.286.5446.1913
78. Reiser J-B, Darnault C, Grégoire C, Mosser T, Mazza G, Kearney A, et al. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nat Immunol* (2003) **4**(3):241–7. doi:10.1038/ni891
79. Newell EW, Ely LK, Kruse AC, Reay PA, Rodriguez SN, Lin AE, et al. Structural basis of specificity and cross-reactivity in T cell receptors specific for cytochrome c-I-E(k). *J Immunol* (2011) **186**(10):5823–32. doi:10.4049/jimmunol.1100197
80. Su LF, Kidd BA, Han A, Kotzin JJ, Davis MM. Virus-specific CD4(+) memory-phenotype T cells are abundant in unexposed adults. *Immunity* (2013) **38**(2):373–83. doi:10.1016/j.immuni.2012.10.021
81. Heath WR, Miller JF. Expression of two alpha chains on the surface of T cells in T cell receptor transgenic mice. *J Exp Med* (1993) **178**(5):1807–11. doi:10.1084/jem.178.5.1807
82. Padovan E, Casorati G, Dellabona P, Meyer S, Brockhaus M, Lanzavecchia A. Expression of two T cell receptor alpha chains: dual receptor T cells. *Science* (1993) **262**(5132):422–4. doi:10.1126/science.8211163
83. Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, Shortman K. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *J Exp Med* (1993) **178**(2):615–22. doi:10.1084/jem.178.2.615
84. Selin LK, Varga SM, Wong IC, Welsh RM. Protective heterologous antiviral immunity and enhanced immunopathogenesis mediated by memory T cell populations. *J Exp Med* (1998) **188**(9):1705–15. doi:10.1084/jem.188.9.1705
85. Yang HY, Joris I, Majno G, Welsh RM. Necrosis of adipose tissue induced by sequential infections with unrelated viruses. *Am J Pathol* (1985) **120**(2):173–7.
86. Chen HD, Fraire AE, Joris I, Brehm MA, Welsh RM, Selin LK. Memory CD8+ T cells in heterologous antiviral immunity and immunopathology in the lung. *Nat Immunol* (2001) **2**(11):1067–76. doi:10.1038/ni727
87. Mason D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol Today* (1998) **19**(9):395–404. doi:10.1016/S0167-5699(98)01299-7
88. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol* (2012) **12**(9):669–77. doi:10.1038/nri3279
89. Jenkins MK, Chu HH, McLachlan JB, Moon JJ. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. *Annu Rev Immunol* (2010) **28**:275–94. doi:10.1146/annurev-immunol-030409-101253
90. Murali-Krishna K, Ahmed R. Cutting edge: naive T cells masquerading as memory cells. *J Immunol* (2000) **165**(4):1733–7.
91. Ernst B, Lee DS, Chang JM, Sprent J, Surh CD. The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. *Immunity* (1999) **11**(2):173–81. doi:10.1016/S1074-7613(00)80092-8
92. Goldrath AW, Bevan MJ. Low-affinity ligands for the TCR drive proliferation of mature CD8+ T cells in lymphopenic hosts. *Immunity* (1999) **11**(2):183–90. doi:10.1016/S1074-7613(00)80093-X
93. Kieper WC, Jameson SC. Homeostatic expansion and phenotypic conversion of naïve T cells in response to self peptide/MHC ligands. *Proc Natl Acad Sci U S A* (1999) **96**(23):13306–11. doi:10.1073/pnas.96.23.13306
94. Muranski P, Chmielowski B, Ignatowicz L. Mature CD4+ T cells perceive a positively selecting class II MHC/peptide complex in the periphery. *J Immunol* (2000) **164**(6):3087–94.
95. Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, Dorrell L, et al. T cell cross-reactivity and conformational changes during TCR engagement. *J Exp Med* (2004) **200**(11):1455–66. doi:10.1084/jem.20041251
96. Kosmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, et al. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* (2010) **465**(7296):350–4. doi:10.1038/nature08997
97. Cameron BJ, Gerry AB, Dukes J, Harper JV, Kannan V, Bianchi FC, et al. Identification of a titin-derived HLA-A1-presented peptide as a cross-reactive target for engineered MAGE A3-directed T cells. *Sci Transl Med* (2013) **5**(197):197ra103. doi:10.1126/scitranslmed.3006034
98. Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells

in myeloma and melanoma. *Blood* (2013) **122**(6):863–71. doi:10.1182/blood-2013-03-490565

99. Fujinami RS, Oldstone MB. Amino acid homology between the encephalitogenic site of myelin basic protein and virus: mechanism for autoimmunity. *Science* (1985) **230**(4729):1043–5. doi:10.1126/science.2414848
100. Oldstone MB. Molecular mimicry and immune-mediated diseases. *FASEB J* (1998) **12**(13):1255–65.
101. Wucherpfennig KW, Strominger JL. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. *Cell* (1995) **80**(5):695–705. doi:10.1016/0092-8674(95)90348-8
102. Burrows SR, Silins SL, Khanna R, Burrows JM, Rischmueller M, McCluskey J, et al. Cross-reactive memory T cells for Epstein-Barr virus augment the alloresponse to common human leukocyte antigens: degenerate recognition of major histocompatibility complex-bound peptide by T cells and its role in alloreactivity. *Eur J Immunol* (1997) **27**(7):1726–36. doi:10.1002/eji.1830270720

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 September 2013; accepted: 10 December 2013; published online: 26 December 2013.
*Citation: Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN and Antia R (2013) Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. Front. Immunol. **4**:485. doi: 10.3389/fimmu.2013.00485*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Zarnitsyna, Evavold, Schoettle, Blattman and Antia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

GENERAL COMMENTARY
published: 25 December 2013
doi: 10.3389/fimmu.2013.00466

Huge overlap of individual TCR beta repertoires

Mikhail Shugay[1†], Dmitriy A. Bolotin[1†], Ekaterina V. Putintseva[1], Mikhail V. Pogorelyy[1], Ilgar Z. Mamedov[1,2] and Dmitriy M. Chudakov[1,2]*

[1] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia
[2] Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic
*Correspondence: chudakovdm@mail.ru
[†] Mikhail Shugay and Dmitriy A. Bolotin have contributed equally to this work.

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Keywords: adaptive immunity, TCR repertoire, TCR beta, NGS data analysis, overlap

A commentary on

Mother and child T cell receptor repertoires: deep profiling study
by Putintseva EV, Britanova OV, Staroverov DB, Merzlyak EM, Turchaninova MA, Shugay M, Bolotin DA, Pogorelyy MV, Mamedov IZ, Bobrynina V, Maschan M, Lebedev YB and Chudakov DM. Front Immunol (2013) 4:463. doi:10.3389/fimmu.2013.00463

It has been reported that human TCR repertoires commonly carry so-called public clonotypes – CDR3 variants that are often shared between individuals. Cross-comparison of individual immune repertoires has previously revealed the existence of a population of TCR beta CDR3 variants that are identical at the amino acid level for any two donors (1–3). The lower bound for the total overlap between any two given donors' TCR beta repertoires within their CD8+ naïve T cell subset has been estimated as ~14,000 identical amino acid CDR3 variants based on comparison of 200,000–600,000 individual TCR beta clonotypes (1). Here, we have used deep profiling data consisting of 1–2 $\times 10^6$ individual TCR beta clonotypes that we obtained from healthy donors (4) to better estimate the total overlap between TCR beta repertoires for any two individuals.

The apparent paradox is, that the deeper we sequence, the larger is the percentage of observed overlapping clonotypes between the two repertoires, since the number of possible element pairs between the two sets grows geometrically. To demonstrate this, we analyzed TCR beta repertoires for 12 unrelated pairs assembled from a total of nine human donors [adults and children, see Ref. (4) for details]. We plotted the number of identical variants found in samples of increasing size, with up to 10^6 unique CDR3 sequences randomly drawn from the repertoires of each individual in a given pair (**Figure 1**). For every pair, the number of shared clonotypes grew geometrically with the arithmetic growth of the sample size (**Figures 1A–C**, colored lines); at maximum sequencing depth (~1 $\times 10^6$ unique sequences/donor), we observed an average of ~72,000, 68,000, and 6,000 CDR3 variants that were respectively identical at the amino acid, amino acid only/non-nucleotide and nucleotide level. This exceeds previous estimates (1) by several-fold. The greatest overlap was between two donors from whom we obtained ~1 $\times 10^6$ and 1.7 $\times 10^6$ CDR3 variants, where we observed 113,000, 108,000, and 11,000 identical clonotypes at the amino acid, amino acid only/non-nucleotide and nucleotide level, respectively.

The lower bound on total individual TCR beta repertoire diversity has previously been estimated to be 5 $\times 10^6$ unique clonotypes [Ref. (5) and our unpublished data]. With that in mind, we extrapolated our intersection curves by fitting them to a power law model [$Y = aX^b$, as in Ref. (1)], which yielded coefficient "b" close to 2.0 and $R^2 > 0.999$ for all cases (**Figures 1A–C**, dashed lines). We estimated that the total overlap of the TCR beta CDR3 repertoires for two individuals constitutes ~2,200,000, 2,060,000, and 180,000 variants, i.e. 44.1, 41.3, and 3.6% of a given individual's sequence diversity at the amino acid, amino acid only/non-nucleotide, and nucleotide level, respectively.

Thus, the real paradox is that nearly half of the TCR beta CDR3 repertoire is functionally identical between any two individuals, in spite of the fact that the theoretical diversity that can be achieved by TCR beta variants has been estimated to be ~5 $\times 10^{11}$ sequences (1, 6). The results from our extrapolation are direct and evident. We took numerous precautions to exclude contamination in our work, including sequencing of pair-analyzed donor repertoires in separate Illumina lanes (4). Even if contaminations were present, these would not affect overlap at the amino acid only/non-nucleotide level (**Figure 1B**). Furthermore, we performed CDR3 extraction and error correction with MiTCR (http://mitcr.milaboratory.com/) using the stringent ETE algorithm, which eliminates 98% of PCR and sequencing errors with minimal loss of natural TCR beta diversity (7).

Such large overlap between individuals suggests the existence of a rather limited pool of frequently used functional CDR3 sequences. To further investigate this, we calculated the lower and upper bounds of the Chao richness estimate as described in Ref. (8) based on the numbers of singletons and doubletons (sequences observed in one and two individuals, respectively) in 12 paired donors' samples. From this model, we obtained a confidence interval of 1.2 $\times 10^7$ to 5.4 $\times 10^7$ unique amino acid CDR3 sequences, at a significance level of $\alpha = 0.001$.

These findings represent a shift in our understanding of human adaptive immunity. It now appears likely that recombinatorial biases (3, 9) and thymic selection (4, 10, 11) shape our repertoires so

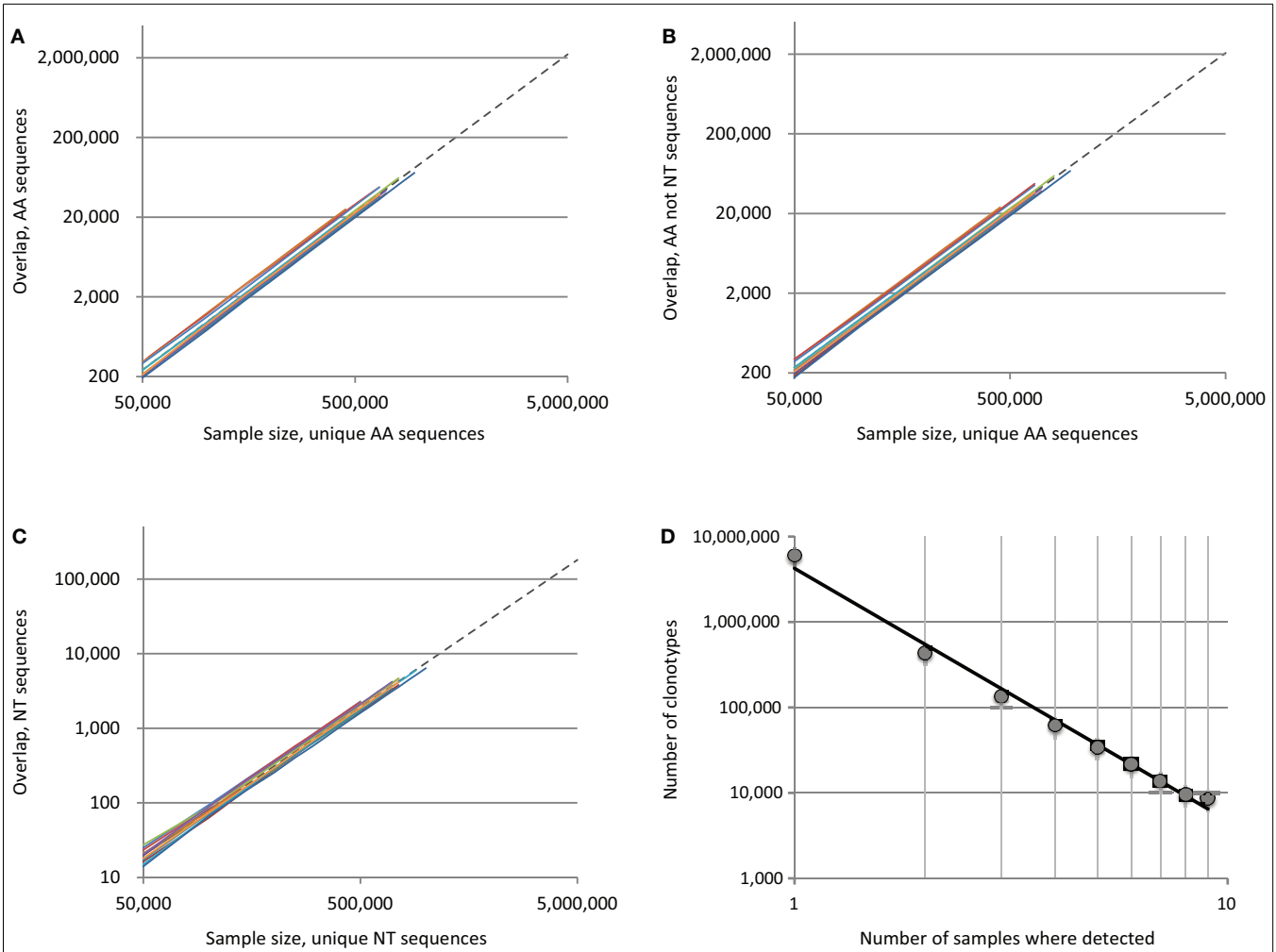


FIGURE 1 | Overlap of individual TCR beta CDR3 repertoires grows geometrically with the number of sequence pairs sampled. Plots indicate the number of shared sequences for 12 unrelated donor pairs in relation to sample size at the level of **(A)** all amino acid sequences, **(B)** amino acid sequence only, excluding matches with identical nucleotide sequences, and **(C)** nucleotide sequences. Each of the 12 colored lines represents the observed overlap between randomly drawn samples of unique CDR3 variants for a different pair of unrelated donors. To extrapolate the predicted level of overlap if the full individual TCR beta repertoires were to be sampled, we plotted fittings of averaged data with a power law ($Y = aX^b$) as dashed lines. **(D)** We plotted the degree to which unique clonotypes were shared among our nine donors, and found that the frequency with which TCR beta clonotypes occur in human repertoires is distributed according to a power law.

tightly that the majority of TCR beta CDR3 variants expressed by naïve T cells leaving the thymus are chosen from a "short-list" of just under 10^8 amino acid variants – even shorter than the 2×10^9 "effective sequence space" estimated by Robins and colleagues (1).

Nevertheless, the repertoire has a complex structure and those clonotypes that are characterized as low-complexity [see figure 7 in Ref. (4)] predominantly form the backbone of the shared clonotype pool. Interestingly, when we examined the intersection of all nine donor samples, we found that the number of donors in which a given clonotype can be detected is distributed according to a power law, with a degree of −2.95 and $R^2 = 0.99$ (**Figure 1D**). These findings confirm the fractal structure of the human TCR beta repertoire that determines the landscape of shared clonotypes (1–3, 12), and may reveal a more complex picture with the deeper profiling experiments.

ACKNOWLEDGMENTS

We are grateful to M. Eisenstein for English editing. This work was supported by the Molecular and Cell Biology program RAS, Russian Foundation for Basic Research (12-04-33139, 12-04-00229, 13-04-00998), and European Regional Development Fund (CZ.1.05/1.1.00/02.0068).

REFERENCES

1. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* (2010) **2**:47ra64. doi:10.1126/scitranslmed.3001442
2. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* (2011) **186**:4285–94. doi:10.4049/jimmunol.1003898

3. Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, et al. Recombinatorial biases and convergent recombination determine interindividual TCRbeta sharing in murine thymocytes. *J Immunol* (2012) **189**:2404–13. doi:10.4049/jimmunol.1102087
4. Putintseva EV, Britanova OV, Staroverov DB, Merzlyak EM, Turchaninova MA, Shugay M, et al. Mother and child T cell receptor repertoires: deep profiling study. *Front Immunol* (2013) **4**:463. doi:10.3389/fimmu.2013.00463
5. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) **114**:4099–107. doi:10.1182/blood-2009-04-217604
6. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* (1988) **334**:395–402. doi:10.1038/334395a0
7. Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, et al. MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* (2013) **10**(9):813–4. doi:10.1038/nmeth.2555
8. Eren MI, Chao A, Hwang WH, Colwell RK. Estimating the richness of a population when the maximum number of classes is fixed: a nonparametric solution to an archaeological problem. *PLoS One* (2012) **7**:e34179. doi:10.1371/journal.pone.0034179
9. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? *Nat Rev Immunol* (2008) **8**:231–8. doi:10.1038/nri2260
10. Turner SJ, Doherty PC, McCluskey J, Rossjohn J. Structural determinants of T-cell receptor bias in immunity. *Nat Rev Immunol* (2006) **6**:883–94. doi:10.1038/nri1977
11. Gras S, Kjer-Nielsen L, Burrows SR, McCluskey J, Rossjohn J. T-cell receptor bias and immunity. *Curr Opin Immunol* (2008) **20**:119–25. doi:10.1016/j.coi.2007.12.001
12. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107

Received: 30 July 2013; accepted: 12 September 2013; published online: 25 December 2013.
*Citation: Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ and Chudakov DM (2013) Huge overlap of individual TCR beta repertoires. Front. Immunol. **4**:466. doi: 10.3389/fimmu.2013.00466*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Shugay, Bolotin, Putintseva, Pogorelyy, Mamedov and Chudakov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 November 2013
doi: 10.3389/fimmu.2013.00379

CD4$^+$ T cell-receptor repertoire diversity is compromised in the spleen but not in the bone marrow of aged mice due to private and sporadic clonal expansions

Eric Shifrut[1], Kuti Baruch[2], Hilah Gal[1], Wilfred Ndifon[1], Aleksandra Deczkowska[2], Michal Schwartz[2] and Nir Friedman[1]*

[1] *Department of Immunology, Weizmann Institute of Science, Rehovot, Israel*
[2] *Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
António Gil Castro, University of Minho, Portugal
Maria L. Toribio, Spanish Research Council (CSIC), Spain

****Correspondence:***
Nir Friedman, Department of Immunology, Weizmann Institute of Science, 76100 Rehovot, Israel
e-mail: nir.friedman@weizmann.ac.il

Reduction in T cell receptor (TCR) diversity in old age is considered as a major cause for immune complications in the elderly population. Here, we explored the consequences of aging on the TCR repertoire in mice using high-throughput sequencing (TCR-seq). We mapped the TCRβ repertoire of CD4+ T cells isolated from bone marrow (BM) and spleen of young and old mice. We found that TCRβ diversity is reduced in spleens of aged mice but not in their BM. Splenic CD4+ T cells were also skewed toward an effector memory phenotype in old mice, while BM cells preserved their memory phenotype with age. Analysis of Vβ and Jβ gene usage across samples, as well as comparison of CDR3 length distributions, showed no significant age dependent changes. However, comparison of the frequencies of amino-acid (AA) TCRβ sequences between samples revealed repertoire changes that occurred at a more refined scale. The BM-derived TCRβ repertoire was found to be similar among individual mice regardless of their age. In contrast, the splenic repertoire of old mice was not similar to those of young mice, but showed an increased similarity with the BM repertoire. Each old-mouse had a private set of expanded TCRβ sequences. Interestingly, a fraction of these sequences was found also in the BM of the same individual, sharing the same nucleotide sequence. Together, these findings show that the composition and phenotype of the CD4+ T cell BM repertoire are relatively stable with age, while diversity of the splenic repertoire is severely reduced. This reduction is caused by idiosyncratic expansions of tens to hundreds of T cell clonotypes, which dominate the repertoire of each individual. We suggest that these private and abundant clonotypes are generated by sporadic clonal expansions, some of which correspond to pre-existing BM clonotypes. These organ- and age-specific changes of the TCRβ repertoire have implications for understanding and manipulating age-associated immune decline.

Keywords: TCR repertoire, aging, immune niche, clonal dominance, high throughput sequencing, TCR-seq, CD4+ T cells

INTRODUCTION

Effective T cell immunity is founded on a diverse T cell-receptor (TCR) repertoire. This diversity, generated by the V(D)J recombination mechanism in the thymus (1), is essential for coping with the plethora of invading and fast evolving pathogens. Loss of diversity, whether naturally occurring with age (2) or induced (3), is associated with increased susceptibility to infections, as well as reduced responses to vaccination (4, 5). One of the most dramatic manifestations of aging on the immune system is thymus involution. Toward old age, both in human and mice (6), the thymic epithelial tissue is replaced by connective and adipose tissue (7), causing reduction in *de novo* production of naive T cells through differentiation of precursor cells. Without thymic activity, naïve T cells are thought to be generated only through homeostatic proliferation of existing single-positive T cells (CD4$^+$ and CD8$^+$ T cells). In adult humans, this is the main mechanism for maintenance of the naïve T cell pool, while in mice there is evidence of lingering thymic output of naïve T cells (8).

Although both the phenotypic balance between memory and naïve T cells as well as the ratio of CD4$^+$ to CD8$^+$ T cells do not alter drastically with age (9), this does not indicate that the TCR repertoire is static (10). By using spectratyping to measure the distributions of TCR lengths across Vβ chains, studies have shown that both the CD8$^+$ and CD4$^+$ TCR repertoires in old mice were skewed compared to young mice (11). Moreover, the perturbations in CDR3 lengths were idiosyncratic to each individual. Deviations from the normal distribution for CDR3 lengths are assumed to be caused by massive T cell expansions during aging both in mice and humans (12). These changes in the composition of the TCR repertoire with aging can create vulnerability to pathogens, such as influenza (13), by providing incomplete clonal coverage.

The bone marrow (BM) is considered as the principal immune niche for both CD4$^+$ and CD8$^+$ memory T cells (14). Following

massive clonal expansion during primary immune response against a pathogen, the contraction phase leaves only a small fraction of antigen-specific memory T cells. These long-lived cells reside mainly in the BM and represent the main T cell reservoir for secondary responses. In line with these observations, the BM was proposed as a "nest" for memory T cells (15), which can be expanded by homeostasis-driven proliferation for fighting viral infections (16), tumor (17), and even age-related cognitive loss (18, 19). It was demonstrated that antigen-specific $CD4^+$ T lymphocytes which relocate throughout life to the BM have a slow turnover, but can fast react as professional memory $CD4^+$ T cells, when stimulated (14).

Although effects of aging on TCR diversity have been evaluated in antigen-specific clones (10), it is still unclear what global changes the TCR repertoire undergoes during lifespan. These large-scale changes have been studied mostly using spectratyping (20, 21), a technique that maps the repertoire with very low resolution. Furthermore, the differences between immune niches, such as the BM and spleen (SPL), in terms of their distinct TCR repertoires and their development with age remain to be explored.

Here we show that the TCRβ repertoire of $CD4^+$ T cells is shaped both by their immunological niche and by age. We used high throughput sequencing to map the murine TCRβ repertoire, of both splenic and BM-derived $CD4^+$ T cells. Aged mice display a marked reduction in diversity of splenic T cells, while diversity of BM-derived T cells is relatively constant with age. Moreover, the TCRβ repertoire of splenic T cells in aged mice becomes more similar to the repertoire of BM-derived T cells. The loss of diversity in old mice is associated with expansion of tens to hundreds T cell clones, and occurs in parallel to segregation of the repertoires of different mice, creating distinct and private immune signatures in each aged individual. Finally, we evaluated clonal expansion and convergent recombination (22) in aging in order to find evidence for the mechanism that creates private repertoires in old age. We show multiple occurrences of sharing at the nucleotide (nt) level between TCR sequences derived from BM T cells and from massively expanded SPL T cell clones of the same aged animal. These results suggest that the degenerate repertoire in old age is shaped by rare events of massive clonal expansions which allow distinctive T cell clones to dominate the immune repertoire of individuals.

MATERIALS AND METHODS

ANIMALS

Inbred male 6- to 8-week-old C57BL/6 mice were supplied by the Animal Breeding Center of The Weizmann Institute of Science. Inbred male 17- to 20-months-old C57BL/6 mice were supplied by the National Institute on Aging (NIA). Aged mice were allowed 1 month adaptation period following shipment from the NIA to our laboratory. All animals were handled according to regulations formulated by The Weizmann Institute's Animal Care and Use Committee and maintained in a pathogen-free environment.

SAMPLE PREPARATION AND CD4+ T CELLS ISOLATION

Prior to tissue collection, mice were intracardially perfused with PBS. Spleens were mashed with a syringe plunger and treated with ammonium-chloride potassium (ACK) lysing buffer to remove erythrocytes. BM was extracted from the femur and tibiae of the mice. Single-cell suspensions of the samples were loaded on MACS column (Miltenyi Biotec) and $CD4^+$ T cells were isolated according to manufacturer's protocol.

FLOW CYTOMETRY AND ANALYSIS

The following fluorochrome-labeled mAbs were used according to the manufacturers' protocols: PercpCy5.5-conjugated anti-TCRβ, PE-conjugated anti-CD4, FITC-conjugated anti-CD44, and APC-conjugated anti-CD62L (BD Pharmingen and eBioscience). Cells were analyzed on an LSRII cytometer (BD Biosciences) using FACSDiva (BD Biosciences) and FlowJo (Tree Star) softwares. In each experiment, relevant negative-control groups and single-stained samples for each tissue were used to identify the populations of interest and to exclude others.

LIBRARY PREPARATION FOR TCR-SEQ

All libraries in this work were prepared and pre-processed as published (23). Briefly, we extracted total RNA from $CD4^+$ T cells (from spleen or BM) of C57BL/6 mice using RNeasy Mini Kit (Qiagen). The RNA was reverse transcribed using SuperScript II reverse transcriptase (Invitrogen) and a TCR Cβ-specific primer linked to the 3′-end Illumina sequencing adapter. The resulting cDNA was then amplified using PCR (Phusion; Finnzymes) with a Cβ-3′adp primer and a set of 23 Vβ-specific 5′ primers, each of which was anchored to a restriction site sequence for the ACUI restriction enzyme. PCR products were then cleaned using QIAquick PCR purification kit (Qiagen), followed by enzymatic digestion with ACUI (New England BioLabs). The ACUI enzyme was used to cleave the amplicons such that sequencing starts closer to the V-D junction region. This allows for good coverage of CDR3β with a single Illumina read. This was followed by ligation of a 5′ Illumina adaptor (T4 ligase; Fermentas), which also contained a 3-nt tag for sample multiplexing. A second round of PCR amplification was performed, using primers for the 5′ and 3′ Illumina adapters. Final PCR products were run on a 2% agarose gel, cut at the desired length, and purified using Wizard SV Gel and PCR Clean-Up System (Promega) to produce the final library. The libraries were sequenced using Genome Analyzer II (Illumina).

PRE-PROCESSING AND ERROR CORRECTION FOR RAW READS

We filtered out raw reads containing bases with Q-value ≤ 30, and then separated the remaining reads according to their barcodes. Then, we aligned the reads to each of the germline Vβ/Jβ gene segments from IMGT (24) using the Smith–Waterman algorithm. Each read was assigned its best-aligning Vβ/Jβ if the number of matching nt (alignment length) was above a threshold: 11 nt for Vβ, 9 nt for Jβ. To reduce the effect of sequencing errors, we clustered (hierarchically) reads assigned the same Vβ and Jβ genes to correct up to 2 nt misincorporation errors. Then, we annotated the sequences by matching the Dβ to the junction, identifying deleted/inserted nt and elongated the read to its full CDR3β length (by IMGT convention). Finally, we translated the nt sequences into amino acid (AA) CDR3β. For the entire analysis here, we used only sequences that are fully annotated (V, J segments assigned), are in-frame (i.e., they encode for a functional peptide, without stop codons), have a cluster size of at least two and have less than 2 bp enzyme cleavage error. We also corrected the copy-number, to adjust for PCR and sub-sampling bias, as published (23).

STATISTICAL ANALYSIS

All statistical analysis was performed using R Statistical Software (R (25)). We also used ShortRead package (26) for the pre-processing pipeline, "ineq" package (27) to calculate the Gini coefficient and "ggplot2" (28) for generating figures. Statistical tests performed are stated in the text.

RESULTS

DIVERSITY OF THE SPLENIC TCRβ REPERTOIRE IS COMPROMISED IN OLD MICE

We aimed to explore the changes in the repertoire landscape at old age, with emphasis on evaluating the diversity of the TCR repertoire. To accomplish this, we measured using TCR-seq the TCRβ repertoire in mice from two age groups: 6–8 weeks old (termed "young," $n = 3$) and 17–20 months old ("old," $n = 3$). In addition, to evaluate the differences between immune organs, we isolated $CD4^+$ T cells from the SPL and BM of each mouse. Properties of all samples are detailed in **Table 1**. On average, we have ∼2e6 sequence reads that have passed the quality threshold (see Materials and Methods), for each sample. These quality-filtered reads produced an average of ∼2.8e5 reads that could be annotated with full CDR3β sequence properties (see Materials and Methods), including translation to an in-frame AA sequence. BM samples resulted in about 10-fold less annotated reads compared with SPL samples. In total, we have found 108,124 distinct CDR3β AA sequences in these 12 samples.

Table 1 | Sample properties.

Sample name	Age group	Organ	Raw reads	Total annotated	Unique reads
ySP1	Young	Spleen	2,653,822	1,075,670	77,991
ySP2	Young	Spleen	804,529	134,372	12,298
ySP3	Young	Spleen	1,426,160	470,363	32,293
yBM1	Young	BM	1,538,612	84,111	6,167
yBM2	Young	BM	1,211,410	4,492	722
yBM3	Young	BM	2,857,600	65,879	6,183
oSP1	Old	Spleen	4,325,524	459,288	18,363
oSP2	Old	Spleen	2,790,495	384,664	17,004
oSP3	Old	Spleen	2,226,719	579,364	9,594
oBM1	Old	BM	1,747,231	16,653	1,443
oBM2	Old	BM	2,036,502	87,758	8,964
oBM3	Old	BM	1,533,444	52,905	4,076

To evaluate the diversity of TCR sequences in the samples, we first checked the cumulative frequencies of clonotypes, ordered by their rank. Hence, we sorted all of the AA clonotypes by their frequency in ascending order. To adjust for varying sample size, we normalized the rank between 0 (the rarest clone) to 1 (the most abundant clone). We then calculated the mean cumulative frequency for increasing rank bins across mice belonging to the same group (**Figure 1A**). In this representation, also known as a Lorenz curve, a repertoire that has maximal diversity (i.e., all clonotypes are present in equal frequency), will be plotted as a straight line across the diagonal. In contrast, a skewed repertoire (i.e., few and very abundant clonotypes dominate the sample) will deviate below the diagonal with a sharp incline only toward the higher ranks. We observe that the most skewed repertoire belongs to the old SPL, while BM from young mice is the most diverse of the repertoires studied here. The old SPL repertoire had a decreased diversity compared with that of the young SPL, while the BM repertoire showed only a slight decrease in diversity with age.

As another measure for repertoire skewness, we applied the Gini coefficient for inequality, used to measure evenness of wealth distribution in economics, which was applied recently for evaluation of TCR repertoire diversity (29). High values of the Gini coefficient, which ranges from 0 to 1, are indicative of a skewed repertoire. We calculated the Gini coefficient for each of the samples and grouped the results by organ and age (**Figure 1B**). The Gini coefficient is highest for the old SPL group, consistent with the Lorenz curve. The decline in clonal equality with age is evident in the spleen ($p < 0.05$, Student's t test), but not in the BM ($p = 0.42$).

We used an additional metric for measuring diversity in TCR samples, the Simpson's diversity index (30, 31). This metric takes into account both the number of unique clonotypes and their relative frequency. The Simpson's diversity index represents the probability that any two clonotypes randomly drawn from the sample will have different sequences. The Simpson index ranges from 0 to 1, with 1 representing maximal diversity, i.e., all clonotypes are present in equal sizes. For each sample, we calculated the mean Simpson's diversity index for 500 randomly sampled clones in 1,000 iterations (**Figure 1C**). Consistent with our observation for the skewness of the repertoire in old mice, the old SPL group has a significantly lower diversity compared with the young SPL [$p < 0.001$, permutations test, see Ref. (30)]. To conclude, we find that the diversity of the splenic TCRβ repertoire is significantly reduced at old age, based on the three analysis methods. Reduction in the diversity of the BM repertoire is minimal and not statistically significant with current sample sizes.

SPLENIC $CD4^+$ T CELLS ARE SKEWED TOWARD AN EFFECTOR MEMORY PHENOTYPE

Next, we wished to determine whether these two immunological compartments, SPL and BM, age differently in terms of the memory phenotype of $CD4^+$ T cells. Using flow cytometry, we measured the proportions of effector memory (T_{EM}) and central memory (T_{CM}) phenotypes in the $CD4^+$ memory T cell compartment (**Figure 1D**). We found that T_{EM} $CD4^+$ T cells are significantly more abundant in spleens of old mice ($93.5 \pm 2.7\%$) compared to young mice ($76.7 \pm 1.8\%$). Increase in the effector phenotype could point to expansion driven by antigen-specific responses and suggests that these $CD4^+$ T_{EM} cells contribute to the skewed repertoire we observe in old spleen. No significant change in T_{EM}/T_{CM} balance was observed in BM samples between age groups (**Figure 1D**), suggesting maintenance of the phenotypic balance in the BM niche.

ORGAN SPECIFIC PATTERNS OF Vβ AND Jβ SEGMENT USAGE ARE CONSTANT WITH AGE

We next examined how the gene segment usage changes with aging (**Figures 2A,B**). Qualitatively, the Jβ distributions look similar in all samples, with few segments that differ in frequency between spleen and BM. For example, Jβ1.3 is over-expressed

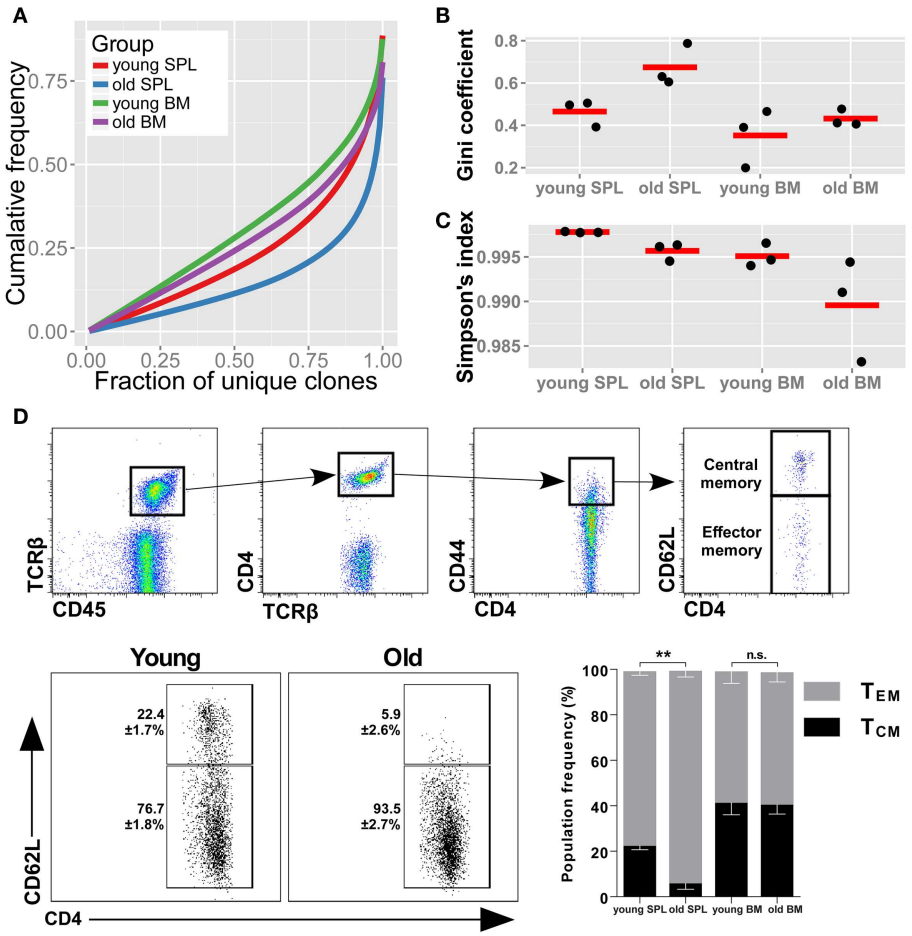


FIGURE 1 | TCRβ repertoire is less diverse in the spleens of old mice. **(A)** Skewness of the TCRβ repertoire for CD4+ T cells from spleen and BM of young and aged mice. For each mouse, clonotypes were ordered by frequency. We then compared the cumulative frequency at each rank (normalized to sample size). From the curves, which represent the mean for each group, we observe that old SPL has the most skewed repertoire. **(B)** Gini coefficient per group. Horizontal lines represent the mean for each group and dots are individual samples. Old SPL group has the highest values, thus it is the most skewed. **(C)** Simpson's diversity index calculated for each mouse (dots). Horizontal lines represent the mean for each group. The old SPL group has a significantly lower diversity than the young SPL group. **(D)** Phenotypic changes of CD4+ memory T cells in aging. Top panels show flow cytometric gating strategy. Old spleen samples have significantly higher percentage of effector memory T cells compared to young spleen samples (bottom panels and bar graph). BM samples have similar central/effector memory ratios across age groups (mean ± SE of each group ($n = 4$–5 per group; $^{**}P < 0.01$; Student's t test).

in SPL samples (both young and old) compared to BM samples, whereas Jβ2.7 is under-expressed in SPL samples. The Vβ distributions vary between samples to a larger extent, evidenced also by the overall lower correlation scores (**Figure 2C**). Inter-group correlations in gene segment usage (**Figure 2D**) show similarity between organ specific repertoires across age, which is higher than between repertoires of different organs within age groups. Thus, gene segment usage is more similar between young and old BM samples, and between young and old SPL samples; it is less similar between BM and SPL samples, both in young and aged mice. This suggests that the tissue microenvironment plays a major factor in shaping of the TCRβ repertoire.

Spectratyping analysis is often used to test for skewness and clonal dominance in TCR repertoires. Thus, we generated virtual spectratypes of CDR3 length distributions for sequences grouped by their Vβ segment. Specifically, as our analysis detected skewness in old SPL samples, we aimed to test if a particular CDR3 length was dominant in that age group (see **Figure 2E** for representative plots). This analysis revealed length distributions that are largely homogenous across samples, with neither significant changes in skewness nor enrichment of a specific CDR3 length. Thus, coarse analysis of the $CD4^+$ TCR repertoire, comparing gene segment usage as well as spectratyping analysis, could not explain the measured loss in diversity in aged mice.

BM TCR REPERTOIRES ARE SIMILAR BETWEEN MICE AND AGE GROUPS, WHILE SPL TCR REPERTOIRES CHANGE AND BECOME PRIVATE WITH AGE

TCR-seq allows for comparison of repertoires at a higher resolution, beyond gene segment usage or CDR3 length distributions,

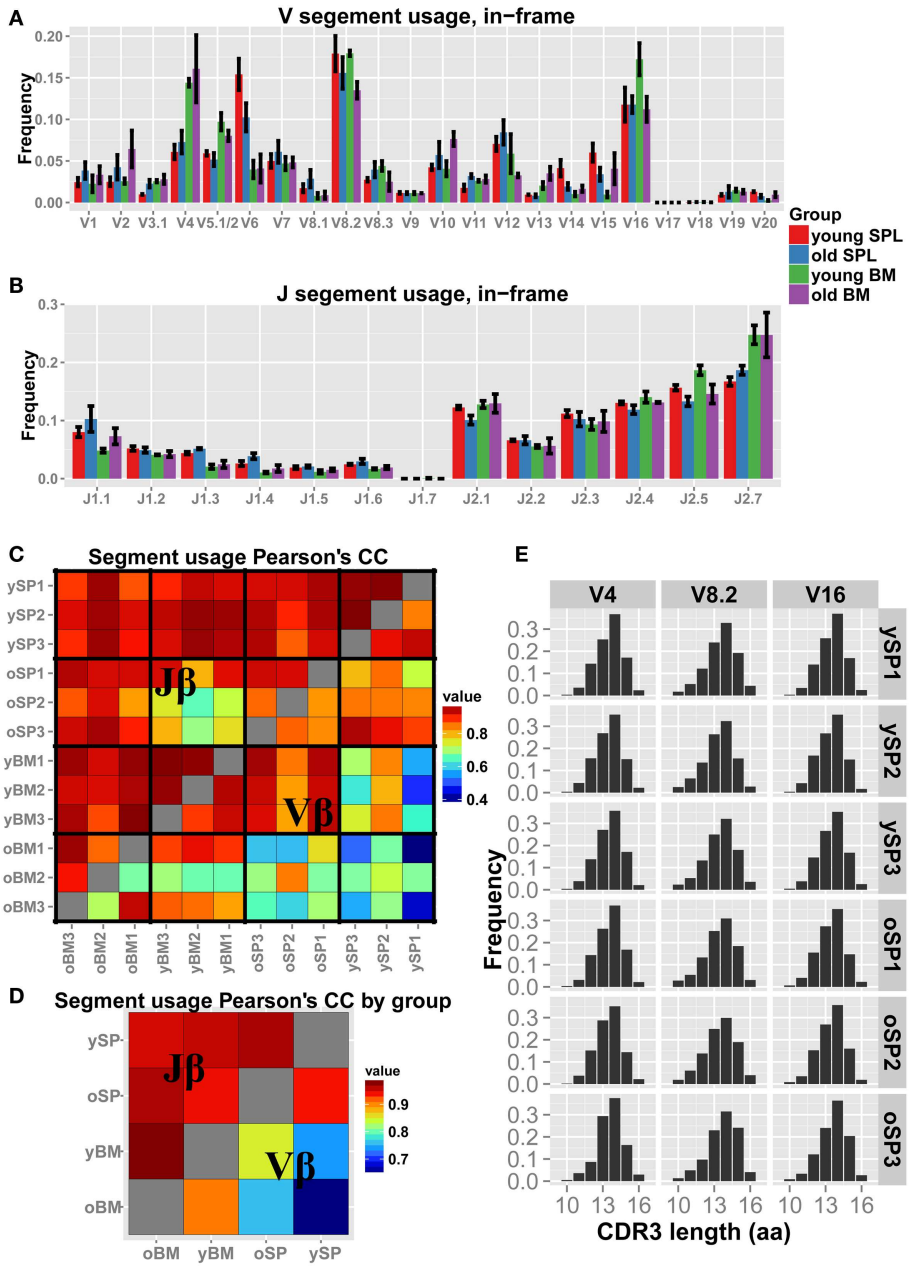


FIGURE 2 | Vβ and Jβ usage in aged mice. **(A,B)** Each bar represents the mean frequency of a gene segment in that group of mice. Error bars are SEM. **(C)** Pairwise correlations of Vβ and Jβ usage between all pairs of mice. We observe higher correlations between mice in Jβ usage (upper triangle) compared to Vβ usage (lower triangle). **(D)** Correlation of the gene segment usage between all pairs of mice averaged over groups. We detect a general high correlation in the data, with inter-tissue similarity across age groups. BM = bone marrow, SPL = spleen, $n = 3$ for all groups. **(E)** "Virtual" spectratypes. Each bar represents the relative frequency of a particular CDR3 length (in amino acids) for three representative Vβ segments, stated above each panel, measured across individual SPL samples. No significant changes could be detected in CDR3 length distributions across age groups.

by analysis at the level of individual TCR sequences. Hence, we assessed similarity between samples by comparing frequencies of overlapping clonotypes. This comparison is more stringent than measuring Vβ/Jβ usage, as we search for the same exact AA clonotypes and compare their observed frequency in each pair of samples. We find that, in accordance to our findings for the Vβ/Jβ usage, all BM samples, regardless of age, display a high similarity in the frequencies of shared AA clonotypes (**Figure 3A**). We also notice that the old SPL group is non-homogenous, i.e., the inter-sample similarity (between individuals) is lower when compared to that found in the young SPL group. To further illustrate this point, we calculated the mean of all pairwise correlation coefficients for each group comparison (e.g., young SPL vs. old SPL, young SPL vs. old BM, etc.). This analysis estimates the overall

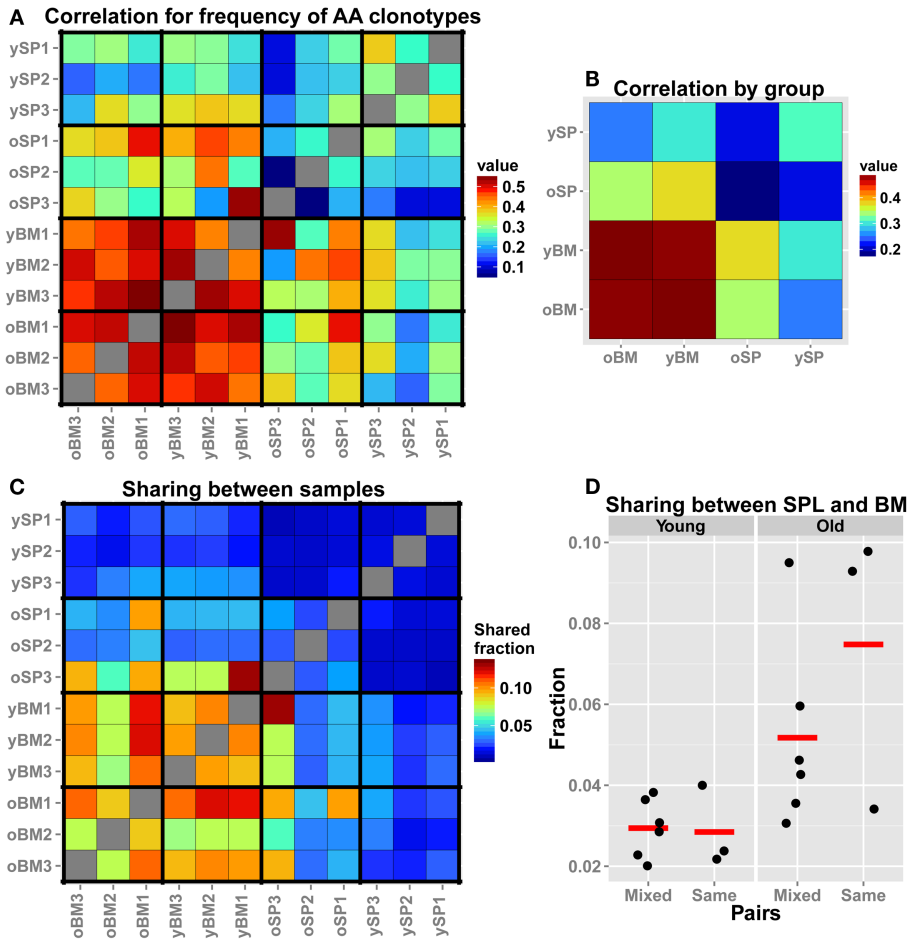


FIGURE 3 | Comparison of the TCR repertoires in different tissues and age groups at the level of AA clonotypes. **(A)** For each pair of mice, we calculated the Spearman's correlation for frequencies of clonotypes that are shared by that pair. As the frequency distribution of TCR repertoire is over-dispersed, we used the log-transformed values as input for the rank correlation test. The BM groups have high similarity between the samples, and also between age groups. **(B)** For each group of mice ($n = 3$), we averaged the correlation scores of **(A)**. Scores along the diagonal indicate the intra-group similarity. Again, the homogeneity in the BM samples within groups and across age is evident. In addition, the repertoire of the old SPL group has a higher correlation with the young BM and old BM groups, and a low intra-group correlation. **(C)** Sharing of clones between samples. From each sample, we randomly chose 300 AA clones and calculated pairwise sharing. The values represent the fraction of the sample that is shared between any particular pair, averaged over 100 iterations. BM repertoires show high level of sharing between individual mice and across age groups. **(D)** A comparison of sharing between BM and SPL repertoires within the same animal or from different animals (Mixed).

similarity between groups (off-diagonal elements in the matrix of **Figure 3B**) and also between samples from the same group (diagonal elements, **Figure 3B**). We observe that the old SPL group has the lowest within-group correlation, and that it is similar to some degree to the young BM and old BM groups, but not to the young SPL group. This suggests that, in aged mice, some clones that were resident in the BM at younger age have migrated to the spleen. BM samples are similar both between and within age groups (**Figure 3B**).

As another measure for similarity, we evaluated the number of overlapping sequences between pairs of samples. In order to control for different sample sizes (**Table 1**), we randomly sampled a collection of 300 clonotypes from each sample and tested how many of these clonotypes are shared between all possible pairs of samples. We iterated this test 100 times to reduce sampling noise and calculated the mean for each pair of samples (**Figure 3C**). We found that the fraction of shared clones within the BM samples is higher than within SPL samples of both age groups. Average sharing of 10% was found between the young BM samples and 9% between the old BM samples. Sharing between SPL samples was much lower (1% for young SPL samples and 3% for old SPL samples). Moreover, 10% of clones are shared on average between young and old BM samples, whereas only 1% are shared between young and old SPL samples. This supports our previous results showing that aging of the immune system affects the composition of the SPL repertoire, but has little influence on the repertoire of BM-resident $CD4^+$ T cells.

We next focused on the inter-tissue sharing of clones by comparing the overlap between SPL and BM repertoires from the same

animal, to the overlap observed between SPL and BM repertoires taken from different animals (**Figure 3D**). In general, there are more clones shared across niches in old animals compared to young animals. Interestingly, we notice that in two out of three old mice, more clones are shared between SPL and BM repertoires that are derived from the same animal, compared with SPL and BM repertoires that are derived from different animals. However, we do not observe this pattern in young animals. These results suggest that with age, a private set of clones is expanded in each individual, contributing to the reduction in SPL repertoire diversity. Furthermore, as the overlap between the repertoires of the spleen and BM niches increases in old age, the repertoire of the whole animal becomes less diverse and degenerate.

CLONAL DOMINANCE IS PREVALENT IN SPLENIC CD4$^+$ T CELLS FROM OLD MICE

Following our observation that the repertoire of splenic T cells from old mice becomes less diverse, private and more similar to the BM repertoire, we next focused on analysis of properties of specific clones that contribute to these phenomena. To that end, we first pooled the top 300 AA clonotypes from each sample to a unified list of 2,108 unique AA sequences. Then, we clustered the log-transformed frequencies for all the sequences in the unified list across all samples (**Figure 4A**). We observe that young SPL samples share many of these top clonotypes, indicating a baseline similarity in the repertoire of young mice. In contrast, the old SPL samples are distinct from each other and each individual presents a unique subset of highly abundant clonotypes, which have intermediate to low frequencies in the young SPL. Furthermore, BM samples from old mice share several abundant clones with their paired SPL samples, consistent with the intra-mouse sharing we observed above (**Figure 3D**).

To reveal if the sharing of AA clonotypes in the old mice samples is also present at the nt level, we picked three representative AA clonotypes that are shared between the SPL and BM samples of old mice (**Figure 4B**). Strikingly, we found that in all three cases the same nt sequence encodes the AA clonotype that is highly frequent in both the SPL and BM. This is a strong indication that the event of clonal expansion occurred for a particular T cell clone, causing clonal dominance in aged mice, which is evident in both BM and SPL of the animal. In contrast, AA clonotypes that are highly frequent in young SPL samples typically show high convergent recombination where many nt sequences encode for the same AA clonotype (**Figure 4B**).

Following this observation, we extended this analysis and counted the number of nt clones that are shared between BM and SPL of the same animal, which are not shared by any other sample (SPL or BM) from other animals. We find that more nt sequences are shared exclusively between the BM and SPL of the same animal in old mice, but not in young mice (**Figure 4C**). This reflects again the private repertoires generated in old age, in parallel with the increased similarity between the BM and SPL repertoires.

These results suggest that expanded clonotypes in aged mice show clonal dominance, that is, the same TCR nt sequence is responsible for most observed TCRs that have the same AA sequence. To test this hypothesis, we directly calculated clonal dominance in old mice compared to young mice. For each mouse, we considered only the 300 most abundant AA clonotypes that are encoded by at least two distinct nt sequences. Then, we calculated the ratio between frequencies of the most abundant nt sequence and the least abundant nt sequence encoding each AA clonotype (**Figure 4D**). In the old SPL group, there is over a 100-fold difference on average, between the maximal frequency and the minimal frequency of nt sequences coding for the same AA sequence. This ratio, R, is significantly higher in old SPL samples ($R = 116$) than in young SPL ($R = 36$). The ratio in the BM is low for both age groups ($R = 13$ for young BM and $R = 16$ for old BM). This supports the hypothesis that expanded clonotypes in the old SPL represent events of massive clonal expansion of a particular T cell clone.

Finally, to illustrate the global changes that the repertoire undergoes with aging, we plot the number of unique nt sequences encoding for the same AA clonotype (convergent recombination level) against the frequency of that clonotype, for all AA clonotypes from all spleen samples (**Figure 4E**). We notice that the old SPL group contains a subset of clonotypes that are highly expanded and have low to moderate convergent recombination (encoded by up to 10 different nt sequences). Clonotypes with similar properties are not found in the young SPL group. This supports the hypothesis that sporadic clonal expansion is a major factor in shaping the repertoire in old mice. Also, there are very few clonotypes ($n = 3$, 0.007% of the clonotypes) with convergent recombination higher than 10 in the old SPL group, whereas in the young SPL group there are many such clonotypes ($n = 152$, 0.16%).

DISCUSSION

The immune system undergoes changes with aging, contributing to an overall increase in neurodegenerative diseases and decrease in autoimmune inflammatory diseases. In addition, susceptibility to infectious diseases inclines with age (32) due to a combination of several factors such as immune senescence (33), transcriptional changes (34), and loss of *de novo* production of naïve T cells (35). Aged individuals are particularly vulnerable to newly encountered pathogens, as TCR diversity is severely diminished in old age. Here we explored the consequences of aging on the TCR repertoire in mice, with focus on the underlying causes for loss of diversity.

We applied TCR-seq on CD4$^+$ T cells isolated from the BM and spleen of young and aged mice. First, our focus was on measuring diversity across all samples. As observed before for CD8$^+$ T cells (20), we found that the diversity of the splenic CD4$^+$ T cell compartment also declines in aged mice. Reduced diversity was revealed both by high Gini inequality coefficient and a low Simpson diversity index. However, in BM samples only a minor reduction of diversity was detected with aging, which was not statistically significant. Examining the memory phenotype of the CD4$^+$ T cells in these two niches revealed a similar pattern; while in the spleen the memory phenotype of the cells strongly shifted toward effector memory during aging, the proportions of effector and central memory T cells in the BM remained constant. This can be attributed to the nature of the BM immune niche as an immune privileged hematopoiesis site (36), allowing maintenance of only a small subset of T cell clonotypes thus less affected by clonal attrition. Clonal expansion in the BM may be inhibited due to the abundance of quiescence-inducing signals which prevent

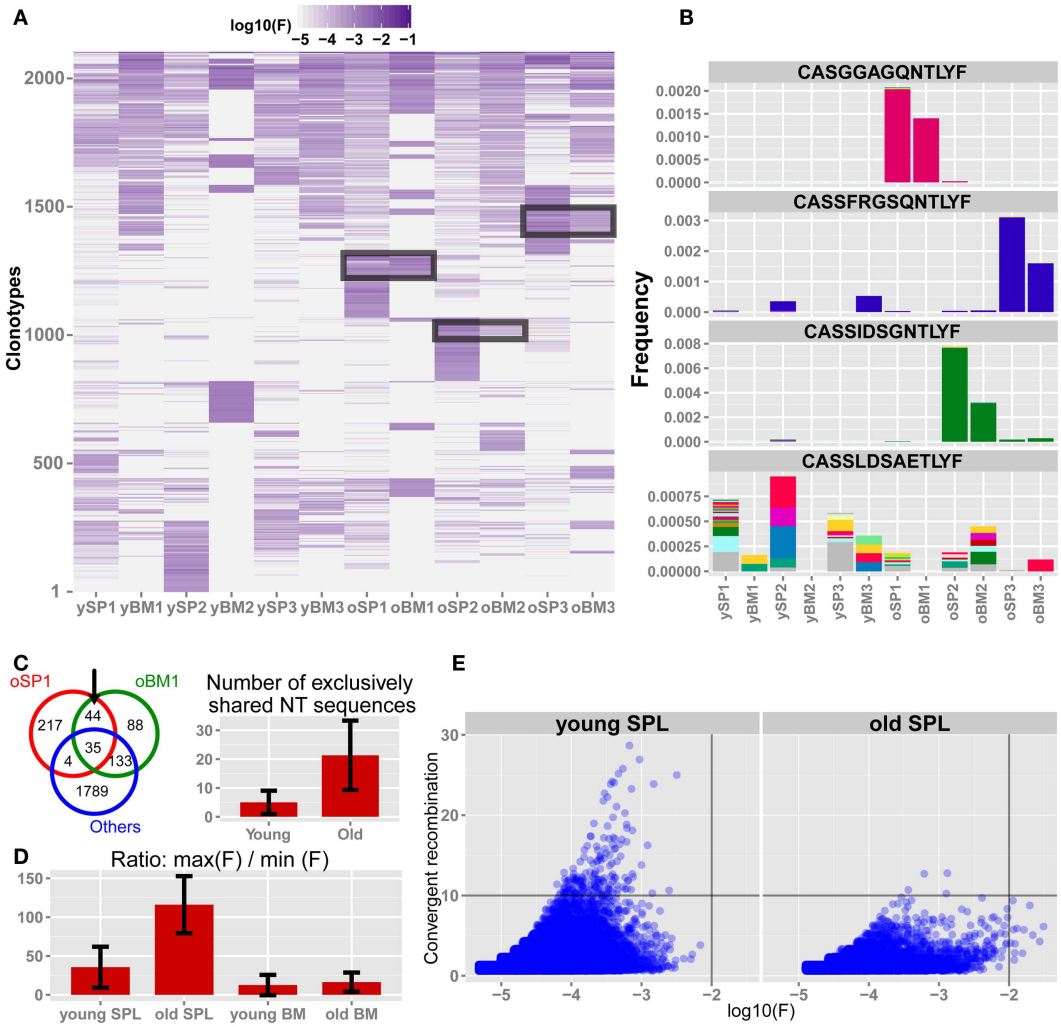


FIGURE 4 | The TCRβ repertoire of old mice is shaped by private clonal expansions. **(A)** Top 300 ranking AA clonotypes from all samples (total of 2,108 clonotypes) were clustered using Euclidian distance. Paired samples from the same animal are adjacent to each other. In old SPL there are subsets of enriched clones, which are unique to each mouse. Also, part of that subset is present in high frequency in the matching BM sample from the same animal (black frames). **(B)** Frequencies of four selected AA clonotypes across all samples. Stacked bars show the nt sequences, uniquely colored, that encode the AA sequence stated above the panel. Top 3 panels show representative clonotypes that are expanded and private in SPL and BM of old mice. Remarkably, these AA sequences are encoded by the same nt sequence in the spleen and BM of these animals. The bottom panel shows a typical abundant AA clonotype in young SPL samples, showing a high level of convergent recombination across most samples in which it is found. **(C)** Sharing of top 300 nt sequences between SPL and BM of the same animal. For each mouse, we calculated the number of exclusively shared nt sequences. The Venn diagram (left) shows an example in which 44 sequences are exclusively shared by BM and SPL of old mouse #1, and are not found in any other sample. Bar plot (right) shows that there are on average more exclusively shared nt sequences within SPL and BM of the same animal in old mice compared to young mice. Error bars indicate standard error. **(D)** Evaluation of clonal dominance. Bars show the average ratio between the frequency of the most frequent and the least frequent nt sequence encoding for the same AA sequence. Values were calculated for the top 300 AA clonotypes from each sample. The ratio is significantly higher in old SPL samples compared to young SPL ($p < 0.05$, Student's t test), suggesting clonal dominance. **(E)** Convergent recombination (# of nt sequences encoding each AA sequence) of AA clonotypes is plotted against their frequency. In old SPL the scatter shows a subset of high frequency clones that are encoded by less than 10 nt sequences (lower-right quadrant). In contrast, in young SPL many clonotypes show high convergent recombination (upper-left quadrant).

extensive proliferation of stem/progenitor cells, found in the BM as a hematopoietic niche (37).

We next evaluated the patterns of gene segment usage in our dataset. Consistent with our previous observation (23), Jβ usage is similar across samples and is not influenced significantly by tissue specificity or age. In general, the highest correlation in Vβ and Jβ usage is observed between the BM samples from both age groups, indicating a relatively stable TCR repertoire in the BM niche across age. However, analysis of distributions of gene segment usage and of CDR3 length do not show significant age-related differences that can explain the observed loss of splenic repertoire diversity. Thus, certain aspects of repertoire dynamics can be evaluated only

with increased resolution, achieved by high throughput methods such as TCR-seq. This may have masked clear detection of decline in repertoire diversity with age in the $CD4^+$ compartment in previous studies that used spectratyping for evaluation of repertoire diversity (21, 38).

Gene segment usage only partially depicts the set of specificities encompassed by the TCR repertoire, thus we focused on the deepest functional level of the repertoire, the CDR3 AA sequence. Here, the similarity between the repertoires of BM niches, even between young and old mice, is emphasized. AA clonotypes in the BM are shared and their frequencies are well-correlated across the samples from different individuals, and also across age groups. This supports the notion of a relatively static composition of clonotypes in the BM niche which maintains the structure of the TCR repertoire of BM-resident $CD4^+$ T cells. In contrast, the SPL samples from aged mice display very distinct repertoires, evidenced by a low intra-group correlation for frequencies of shared AA clonotypes and a low sharing of TCR sequences between mice of this group. This suggests that the loss of diversity we detect in old mice is manifested by private immune responses during lifetime, where in each individual a particular subset of TCR specificities is amplified to dominance. This is in agreement with the pattern observed in antigen-specific response in aged mice (10) but on a more global scale. We observe tens to hundreds of clones that are private and significantly expanded in each old individual's spleen, indicating that decline in repertoire diversity is caused by expansion of a large number of clones through life. Moreover, the repertoire of the old spleen becomes more similar to that of the BM in the aged mice. Of note, in two out of three aged mice, more clonotypes are shared between BM and SPL niches of the same animal compared with sharing between different animals. This trend of exclusive sharing between BM and SPL niche of the same animal is not evident in any of the three young mice. Together with the larger similarity between repertoires of the SPL niche in aged mice to that of the BM, this suggests that specific clones from the BM niche expanded significantly in the periphery, contributing to a skewed, degenerate, and private repertoire in the old SPL. Our phenotypic analysis (**Figure 1D**) suggests that these expanded clones acquire an effector memory phenotype in the old spleen, but more specific analysis is required to validate this hypothesis. In addition, extending our dataset to include additional mice could provide further evidence for the private repertoires in the aged spleen, and for increased similarity between BM and SPL repertoires that we observe in aged mice.

Lastly, we focus on those clones that are common to SPL and BM tissues of aged mice, but exclusive to each animal. We detect clonal dominance in these expanded groups of cells, with the same nt sequence present in high frequency in both niches. The chance of this expansion to occur in two independent events of clonal expansion is highly unlikely. As we find the same exact nt sequence in the BM and SPL of the same animal, we propose that sporadic clonal expansion is the mechanism that shapes the TCR repertoire in aging. This clonal dominance can be realized in the aging immune system, as the "void" created by clonal senescence and exhaustion (39) is more easily filled with rapidly dividing T cell clones. A similar phenomenon was described in other models of similar low grade, chronic sterile innate inflammation, such as obesity, where TCR repertoire is restricted (40). The mechanisms that generate these rare expansions can be response to self-antigens (18), latent infections (32), or driven by accumulating mutations (41).

In summary, we showed that diversity of the splenic $CD4^+$ TCR repertoire declines with age, while the BM repertoire remains largely unchanged. Our results suggest that with age, the TCRβ repertoire of each individual focuses on a certain subset of few hundreds clones out of the potential repertoire, and there is large variability between the subset each individual maintains. This attrition can be explained by a reduction in thymic output of naïve cells with age along with sporadic clonal expansion, which contribute to the clonal dominance we observe in old mice. As a consequence, this phenomenon should be considered when addressing vaccination of the elderly population.

ACKNOWLEDGMENTS

This research was supported by the Minerva Foundation with funding from the Federal German Ministry for Education and Research. Nir Friedman is incumbent of the Pauline Recanati Career Development Chair of Immunology.

REFERENCES

1. Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* (2002) **109**(Suppl):S45–55. doi:10.1016/S0092-8674(02)00675-X
2. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, et al. The influence of age on T cell generation and TCR diversity. *J Immunol* (2005) **174**:7446–52.
3. Nanda NK, Apple R, Sercarz E. Limitations in plasticity of the T-cell receptor repertoire. *Proc Natl Acad Sci U S A* (1991) **88**:9503–7. doi:10.1073/pnas.88.21.9503
4. Kang I, Hong MS, Nolasco H, Park SH, Dan JM, Choi J-Y, et al. Age-associated change in the frequency of memory CD4+ T cells impairs long term CD4+ T cell responses to influenza vaccine. *J Immunol* (2004) **173**:673–81.
5. Wu Y-CB, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front Immunol* (2012) **3**:193. doi:10.3389/fimmu.2012.00193
6. Linton PJ, Dorshkind K. Age-related changes in lymphocyte development and function. *Nat Immunol* (2004) **5**:133–9. doi:10.1038/ni1033
7. Bains I, Thiébaut R, Yates AJ, Callard R. Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output. *J Immunol* (2009) **183**:4329–36. doi:10.4049/jimmunol.0900743
8. Den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, de Boer AB, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* (2012) **36**:288–97. doi:10.1016/j.immuni.2012.02.006
9. Goronzy JJ, Lee W-W, Weyand CM. Aging and T-cell diversity. *Exp Gerontol* (2007) **42**:400–6. doi:10.1016/j.exger.2006.11.016
10. Rudd BD, Venturi V, Li G, Samadder P, Ertelt JM, Way SS, et al. Nonrandom attrition of the naive CD8+ T-cell pool with aging governed by T-cell receptor:pMHC interactions. *Proc Natl Acad Sci U S A* (2011) **108**:13694–9. doi:10.1073/pnas.1107594108
11. Mosley RL, Koker MM, Miller RA. Idiosyncratic alterations of TCR size distributions affecting both CD4 and CD8 T cell subsets in aging mice. *Cell Immunol* (1998) **189**:10–8. doi:10.1006/cimm.1998.1369
12. Schwab R, Szabo P, Manavalan JS, Weksler ME, Posnett DN, Pannetier C, et al. Expanded CD4+ and CD8+ T cell clones in elderly humans. *J Immunol* (1997) **158**:4493–9.
13. Yager EJ, Ahmed M, Lanzer K, Randall TD, Woodland DL, Blackman MA. Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J Exp Med* (2008) **205**:711–23. doi:10.1084/jem.20071140

14. Tokoyoda K, Zehentmeier S, Hegazy AN, Albrecht I, Grün JR, Löhning M, et al. Professional memory CD4+ T lymphocytes preferentially reside and rest in the bone marrow. *Immunity* (2009) **30**:721–30. doi:10.1016/j.immuni.2009.03.015
15. Di Rosa F, Pabst R. The bone marrow: a nest for migratory memory T cells. *Trends Immunol* (2005) **26**:360–6. doi:10.1016/j.it.2005.04.011
16. Slifka MK, Whitmire JK, Ahmed R. Bone marrow contains virus-specific cytotoxic T lymphocytes. *Blood* (1997) **90**:2103–8.
17. Feuerer M, Beckhove P, Bai L, Solomayer EF, Bastert G, Diel IJ, et al. Therapy of human tumors in NOD/SCID mice with patient-derived reactivated memory T cells from bone marrow. *Nat Med* (2001) **7**:452–8. doi:10.1038/86523
18. Baruch K, Ron-Harel N, Gal H, Deczkowska A, Shifrut E, Ndifon W, et al. CNS-specific immunity at the choroid plexus shifts toward destructive Th2 inflammation in brain aging. *Proc Natl Acad Sci U S A* (2013) **110**:2264–9. doi:10.1073/pnas.1211270110
19. Ron-Harel N, Segev Y, Lewitus GM, Cardon M, Ziv Y, Netanely D, et al. Age-dependent spatial memory loss can be partially restored by immune activation. *Rejuvenation Res* (2008) **11**:903–13. doi:10.1089/rej.2008.0755
20. Ahmed M, Lanzer KG, Yager EJ, Adams PS, Johnson LL, Blackman MA. Clonal expansions and loss of receptor diversity in the naive CD8 T cell repertoire of aged mice. *J Immunol* (2009) **182**:784–92.
21. Callahan JE, Kappler JW, Marrack P. Unexpected expansions of CD8-bearing cells in old mice. *J Immunol* (1993) **151**:6657–69.
22. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci U S A* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107
23. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proc Natl Acad Sci U S A* (2012) **109**:15865–70. doi:10.1073/pnas.1203916109
24. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* (2009) **37**:D1006–12. doi:10.1093/nar/gkn838
25. Core Team R. *R: A Language and Environment for Statistical Computing*. Vienna (2013).
26. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* (2009) **25**:2607–8. doi:10.1093/bioinformatics/btp450
27. Zeileis A. *Ineq: Measuring Inequality, Concentration, and Poverty*. (2012).
28. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer (2009).
29. Thomas PG, Handel A, Doherty PC, La Gruta NL. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naive and immune T-cell populations. *Proc Natl Acad Sci U S A* (2013) **110**:1839–44. doi:10.1073/pnas.1222149110
30. Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* (2007) **321**:182–95. doi:10.1016/j.jim.2007.01.019
31. Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol Lett* (2012) **148**:11–22. doi:10.1016/j.imlet.2012.08.002
32. Nikolich-Zugich J. Ageing and life-long maintenance of T-cell subsets in the face of latent persistent infections. *Nat Rev Immunol* (2008) **8**:512–22. doi:10.1038/nri2318
33. Goronzy JJ, Weyand CM. Understanding immunosenescence to improve responses to vaccines. *Nat Immunol* (2013) **14**:428–36. doi:10.1038/ni.2588
34. Chen G, Lustig A, Weng N-P. T cell aging: a review of the transcriptional changes determined from genome-wide analysis. *Front Immunol* (2013) **4**:121. doi:10.3389/fimmu.2013.00121
35. Blackman MA, Woodland DL. The narrowing of the CD8 T cell repertoire in old age. *Curr Opin Immunol* (2011) **23**:537–42. doi:10.1016/j.coi.2011.05.005
36. Fujisaki J, Wu J, Carlson AL, Silberstein L, Putheti P, Larocca R, et al. In vivo imaging of Treg cells providing immune privilege to the haematopoietic stem-cell niche. *Nature* (2011) **474**:216–9. doi:10.1038/nature10160
37. Martinez-Agosto JA, Mikkola HKA, Hartenstein V, Banerjee U. The hematopoietic stem cell and its niche: a comparative view. *Genes Dev* (2007) **21**:3044–60. doi:10.1101/gad.1602607
38. Haynes L, Maue AC. Effects of aging on T cell function. *Curr Opin Immunol* (2009) **21**:414–7. doi:10.1016/j.coi.2009.05.009
39. Almanzar G, Schwaiger S, Jenewein B, Keller M, Herndler-Brandstetter D, Würzner R, et al. Long-term cytomegalovirus infection leads to significant changes in the composition of the CD8+ T-cell repertoire, which may be the basis for an imbalance in the cytokine production profile in elderly persons. *J Virol* (2005) **79**:3675–83. doi:10.1128/JVI.79.6.3675-3683.2005
40. Yang H, Youm Y-H, Vandanmagsar B, Ravussin A, Gimble JM, Greenway F, et al. Obesity increases the production of proinflammatory mediators from adipose tissue T cells and compromises TCR repertoire diversity: implications for systemic inflammation and insulin resistance. *J Immunol* (2010) **185**:1836–45. doi:10.4049/jimmunol.1000021
41. Johnson PLF, Yates AJ, Goronzy JJ, Antia R. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. *Proc Natl Acad Sci U S A* (2012) **109**:21432–7. doi:10.1073/pnas.1209283110

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 August 2013; paper pending published: 23 September 2013; accepted: 24 October 2013; published online: 19 November 2013.
*Citation: Shifrut E, Baruch K, Gal H, Ndifon W, Deczkowska A, Schwartz M and Friedman N (2013) CD4$^+$ T cell-receptor repertoire diversity is compromised in the spleen but not in the bone marrow of aged mice due to private and sporadic clonal expansions. Front. Immunol. **4**:379. doi: 10.3389/fimmu.2013.00379*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Shifrut, Baruch, Gal, Ndifon, Deczkowska, Schwartz and Friedman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 25 December 2013
doi: 10.3389/fimmu.2013.00463

Mother and child T cell receptor repertoires: deep profiling study

Ekaterina V. Putintseva[1], Olga V. Britanova[1], Dmitriy B. Staroverov[1], Ekaterina M. Merzlyak[1], Maria A. Turchaninova[1], Mikhail Shugay[1], Dmitriy A. Bolotin[1], Mikhail V. Pogorelyy[1], Ilgar Z. Mamedov[1], Vlasta Bobrynina[2], Mikhail Maschan[2], Yuri B. Lebedev[1] and Dmitriy M. Chudakov[1,3]*

[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia*
[2] *Federal Scientific Clinical Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia*
[3] *Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic*

Edited by:
Michal Or-Guil, Humboldt University Berlin, Germany

Reviewed by:
Aridaman Pandit, Utrecht University, Netherlands
Nicole Wittenbrink, Humboldt University Berlin, Germany

***Correspondence:**
Dmitriy M. Chudakov, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Miklukho-Maklaya 16/10, Moscow 117997, Russia
e-mail: chudakovdm@mail.ru

The relationship between maternal and child immunity has been actively studied in the context of complications during pregnancy, autoimmune diseases, and haploidentical transplantation of hematopoietic stem cells and solid organs. Here, we have for the first time used high-throughput Illumina HiSeq sequencing to perform deep quantitative profiling of T cell receptor (TCR) repertoires for peripheral blood samples of three mothers and their six children. Advanced technology allowed accurate identification of 5×10^5 to 2×10^6 TCR beta clonotypes per individual. We performed comparative analysis of these TCR repertoires with the aim of revealing characteristic features that distinguish related mother-child pairs, such as relative TCR beta variable segment usage frequency and relative overlap of TCR beta complementarity-determining region 3 (CDR3) repertoires. We show that thymic selection essentially and similarly shapes the initial output of the TCR recombination machinery in both related and unrelated pairs, with minor effect from inherited differences. The achieved depth of TCR profiling also allowed us to test the hypothesis that mature T cells transferred across the placenta during pregnancy can expand and persist as functional microchimeric clones in their new host, using characteristic TCR beta CDR3 variants as clonal identifiers.

Keywords: TCR repertoires, NGS, maternal-fetal exchange, public clonotypes, T cell receptor, haploidentical transplantation, autoimmune diseases, microchimerism

INTRODUCTION

Closeness and relationship between mother and child immunity have been the focus of studies of pregnancy (1), autoimmunity (2–4) and haploidentical transplantations of hematopoietic stem cells (HSCs) (5), and solid organs (6, 7).

In recent years, the potential of next-generation sequencing (NGS) to reveal the full complexity of human and mouse immune receptor repertoires has inspired numerous efforts to develop optimal techniques for achieving large-scale T cell receptor (TCR) and antibody profiling (8–12) and to decipher various aspects of adaptive immunity (8, 9, 11, 13–17). With appropriate library preparation methods (18), NGS techniques now make it possible to perform quantitative analysis of hundreds of thousands or millions of distinct TCR beta complementarity-determining region 3 (CDR3) variants. This individual diversity of TCR beta CDR3 variants, which is generated in the course of V-D-J recombination and the random addition and deletion of nucleotides in the thymus, largely determines the whole diversity of naïve T cells and specificity of T cell immune responses (19, 20).

In the present study, we have used deep NGS profiling to compare TCR beta repertoires of mothers and their children. We achieved a profiling depth of 500,000–2,000,000 unique TCR beta CDR3 clonotypes per donor, and performed comparative analysis with the aim of revealing specific features of TCR repertoires that distinguish related mother-child pairs from unrelated individuals, and how these familial repertoires manifest the influence of inherited factors, such as the elements of TCR recombination machinery and human leukocyte antigens (HLA). By comparing out-of-frame (i.e., non-functional and thus not subjected to selection) and in-frame TCR beta repertoires, we also show the extent of the impact of thymic selection and the common trends in how this process shapes individual repertoires.

Additionally, the profiling depth that we achieved allowed us to look for the potential presence of maternal or fetal microchimeric T cell clones that may have transmigrated through the placenta as mature α/β T cells and which subsequently persist in both related donors, by using characteristic TCR beta CDR3 variants as clonal identifiers.

MATERIALS AND METHODS

SAMPLE COLLECTION

This study was approved by the ethical committee of the Federal Scientific Clinical Center of Pediatric Hematology, Oncology, and Immunology. Blood donors provided informed consent prior to participating in the study. Ten milliliters of peripheral blood samples were obtained from nine systemically healthy Caucasian donors: three mothers (average age 40 ± 4 years) and their six children (average age 11 ± 4 years). Peripheral blood mononuclear

cells (PBMCs) were isolated by Ficoll-Paque (Paneco, Russia) density gradient centrifugation. Total RNA was isolated with Trizol (Invitrogen, USA) in accordance with the manufacturer's protocol.

CONTAMINATION PRECAUTIONS

T cell receptor beta libraries were generated in clean PCR hoods with laminar flow, using reagents of high purity and pipette tips with hydrophobic filters. As an additional precaution, we generated the TCR beta libraries for the two groups being compared – mothers and their children – a month apart, and sequenced the two libraries in two separate Illumina runs to guarantee the absence of inter-library contamination during amplification or on the solid phase of the sequencer.

PREPARING cDNA LIBRARIES FOR QUANTITATIVE TCR BETA PROFILING

cDNA-based library preparation was performed essentially as described previously (9, 12, 16, 18, 21, 22). Briefly, we used the Mint kit (Evrogen, Russia) for first-strand cDNA synthesis. For each donor sample, the whole amount of extracted RNA was used for cDNA synthesis, with 1.5 μg of RNA per 15 μl reaction volume. We incubated the mixture of RNA and priming oligonucleotide BC_R4_short (GTATCTGGAGTCATTGA), which is specific to both variants of the human TCR beta constant (TRBC) segment, at 70°C for 2 min and 42°C for 2 min for annealing. We then added the 5′-adapter for the template switch. The reaction was carried out at 42°C for 2 h, with 5 μl of IP solution added after the first 40 min.

Further cDNA library amplification was performed in two sequential PCRs using Encyclo PCR mix (Evrogen). To capture the maximum number of input cDNA molecules, we used the whole amount of synthesized cDNA for the first PCR amplification. The first PCR totaled 18 cycles with universal primers M1SS (AAGCAGTGGTATCAACGCA) and BC2R (TGCTTCTGATGGCTCAAACAC), which are respectively specific to the 5′-adapter and a nested region of the TRBC segments. The primer annealing temperature was set at 62°C. The products of the first PCR were combined, and a 100-μl aliquot was purified by QIAquick PCR purification Kit (Qiagen) and eluted by 20 μl of EB buffer.

The second PCR amplification was performed for 8–10 cycles with a mix of TCR beta joining (TRBJ)-specific primers and the universal primer M1S ($(N)_{2–4}$(XXXXX)CAGTGGTATCAACGCAGAG), which is specific to the 5′-adapter and is nested relative to the M1SS primer used in the first PCR amplification. XXXXX represents a sample barcode introduced in the second PCR, and $(N)_{2–4}$ are random nucleotides that were added in order to generate diversity for better cluster identification during Illumina sequencing. Primer annealing temperature was set at 62°C.

ILLUMINA HiSeq SEQUENCING

PCR products carrying pre-introduced sample barcodes were mixed together in equal ratio for each of the two groups (mothers and children). Illumina adapters were ligated according to the manufacturer's protocol using NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs, USA). Generated libraries were analyzed using two separate Illumina HiSeq 2000 lanes in separate runs with 100 + 100 nt paired end sequencing using Illumina sequencing primers. Raw sequences deposited in NCBI SRA database (PRJNA229070).

NGS DATA ANALYSIS

TCR beta variable (TRBV) segment identification [using IMGT nomenclature (23)], CDR3 identification (based on the sequence between conserved Cys-104 and Phe-118, inclusive), clonotype clusterization and correction of reverse transcription, PCR, and sequencing errors were performed using our MiTCR software (24)[1]. The sequencing quality threshold of each nucleotide within the CDR3 region was set as Phred >25, with low-quality sequence rescue by mapping to high-quality clonotypes. The strictest "eliminate these errors" correction algorithm was employed to eliminate the maximal number of accumulated PCR and sequencing errors.

STATISTICAL ANALYSIS

We used Jensen–Shannon divergence (JS), which is a symmetrized version of the Kullback–Leibler divergence (KL), to quantify the similarity between the clonotype TRBV gene usage distribution in related and unrelated mother-child pairs. JS and KL are defined as follows (25):

$$\mathrm{JS}(P,Q) = \frac{1}{2}\left(\mathrm{KL}\left(P,\ \frac{P+Q}{2}\right) + \mathrm{KL}\left(Q,\ \frac{P+Q}{2}\right)\right),$$

$$\mathrm{KL}(P,Q) = \sum_i p_i \log_2 \frac{pi}{qi}.$$

Where P and Q correspond to the TRBV gene segment frequency distributions of the two individuals being analyzed, and p_i and q_i stand for the frequency of a particular TRBV gene segment in the first and second individual, correspondingly. For statistical comparison of the JS among related and unrelated mother-child pairs, we used two-tailed, unpaired Student's t-test with P-values <0.05 considered significant. To account for multiple testing, Bonferroni-corrected P-values were used.

We used linear regression to analyze dependency between TRBV-CDR3/CDR3 overlap ratio and the number of shared major histocompatibility complex I (MHC-I) alleles, and calculated the Pearson correlation coefficient. The linear model:

$$\text{TRBV} - \text{CDR3/CDR3 overlap ratio} = b0 + b1 \times [\text{Number of shared MHC alleles}]$$

was fit using the least-squares method. Linear regression and correlation analysis were performed using R programing language[2].

FACS ANALYSIS

We used the following anti-human antibodies: CD3-PC7 (clone UCHT1, eBioscience), CD27-PC5 (clone 1A4CD27, Invitrogen), CD4-PE (clone 13B8.2, Beckman Coulter), CD45RA FITC (eBioscience, clone JS-B3). An aliquot of PBMC was incubated with antibodies for 20 min at room temperature, washed twice with PBS and analyzed via Cytomics FC 500 (Beckman Coulter).

HLA TYPING

The samples were HLA-typed using SSP AllSet Gold HLA-ABC Low Res Kit and SSP AllSet Gold HLA-DRDQ Low Res Kit (Invitrogen) and results were processed using UniMatch software.

[1] http://mitcr.milaboratory.com/
[2] http://www.R-project.org/

RESULTS

We obtained at least 1×10^7 TCR beta CDR3-containing sequencing reads for each mother and about 3×10^6 reads for each child. MiTCR software analysis yielded 500,000–2,000,000 distinct TCR beta CDR3 clonotypes per donor (**Table 1**) – representing a significant portion of the total TCR beta diversity for an individual, which lower bound estimate constitutes ~4 million (8). We then subjected these individual TCR beta datasets to comparative analysis in an effort to identify features that distinguish TCR beta repertoires of related mother-child pairs.

TRBV GENE USAGE

We analyzed the relative usage of TRBV gene segments in mother-child pairs at three levels (see **Figure 1**):

Out-of-frame TCR beta variants

The influence of genetic effects on the recombination machinery, which determines the relative frequencies of TRBV gene segment usage in TCRs generated before selection in the thymus, should be reflected by out-of-frame TCR variants that are not subjected to the pressure of further selective processes. Due to nonsense-mediated decay mechanisms, RNA-based libraries generally contain a low percentage of out-of-frame TCR beta variants (9, 12, 26, 27). Nevertheless, out-of-frame CDR3 sequences constituted ~2.5% of all clonotypes (**Table 1**) – 16,048–45,300 clonotypes per donor – which is sufficiently abundant to perform statistical analysis. These subsets were used to compare TRBV gene segment usage in related and unrelated mother-child pairs before thymic selection.

At this level of out-of-frame non-functional TCR beta variants, Jensen–Shannon divergence in TRBV gene usage was comparable for related and unrelated mother-child pairs, albeit with a non-significant increase in divergence for the latter (**Figure 2A**; **Figures 3A,B**, first 2 bars).

Low-frequency in-frame clonotypes

The pressure of thymic selection can be tracked by comparing TRBV gene segment usage in out-of-frame TCR beta variants relative to those variants represented in naïve T cells. In this work, we did not perform separate TCR profiling of FACS-sorted naïve T cells. We aimed to achieve maximal depth of analysis, and sought to avoid the loss of cells and RNA and general quantitative biases that inevitably arise from the cell sorting process. We estimated the pool of TCR beta clonotypes that predominantly belong to the naïve subset as follows. We used FACS analysis to identify the percentage of naïve $CD27^{high}CD45RA^{high}$ CD3+ T cells for each donor (28). This analysis demonstrated that naïve T cells constitute 40–73% of the T cell population in children and 27–55% of the T cell population in mothers (**Table 1**; **Figure 1**). Since each naïve T cell clone is usually represented by minor numbers of TCR-identical cells in an individual (29), for the purposes of bulk analysis, we hypothesized that the subset of the low-frequency clonotypes that occupies the same share of homeostatic space as the FACS-determined share of naïve T cells for that particular donor (433,293–1,797,650 clonotypes per donor) predominantly includes naïve T cells.

At this level of low-frequency, in-frame TCR beta clonotypes, TRBV gene segment usage was significantly less divergent compared to out-of-frame TCR beta variants, both in related and unrelated pairs (**Figures 2B** and **3**). Additionally, TRBV gene segment usage was significantly more similar for related versus unrelated pairs (**Figures 3A,B**, bars 3, 4). In accordance with JS analysis, comparison within related triplets revealed equalization of the usage of particular TRBV gene segments in low-frequency, in-frame TCR clonotypes compared to out-of-frame TCR variants (**Figure 4**). For example, in each triplet, we saw the usage of TRBV gene segments 12-3, 12-4, 20-1, 21-1, and 23-1 equalize in the low-frequency TCR beta clonotypes pool. We also observed an equalizing decrease in TRBV 7-3 usage in triplets A and C, and an equalizing increase in TRBV 28 usage in triplet B.

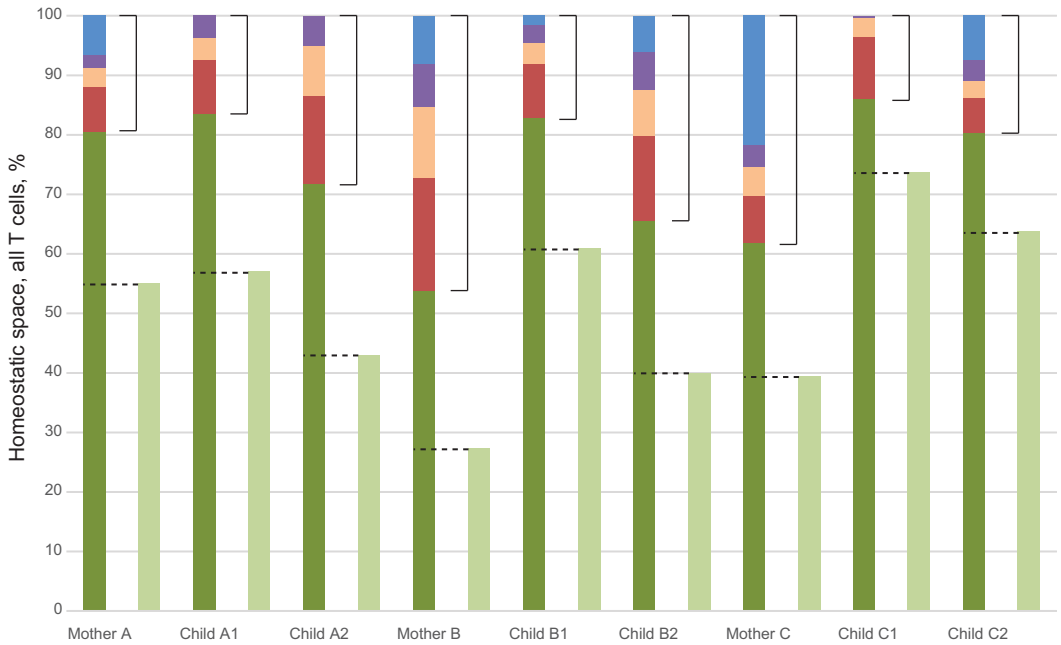
Notably, the observed changes in TRBV gene segments usage were generally similar in different unrelated donors (compare **Figures 4A–C**), and the convergence of TRBV usage after thymic selection (difference of out-of-frame versus in-frame TRBV usage divergence) was not significantly dependent on the number of shared HLA alleles ($R = 0.12$, $P = 0.63$).

High-frequency in-frame clonotypes

The influence of antigen-specific reactions on selection of TRBV gene segments could be tracked by comparing TRBV gene usage in naïve and antigen-experienced T cells. Following the same logic

Table 1 | Sequencing results: TCR beta reads and clonotypes.

Donor	Sex	Age	Paired end sequencing reads	TCR beta CDR3-containing reads (%)	TCR beta clonotypes before error correction	Final TCR beta clonotypes	% Of out-of-frame clonotypes of all clonotypes	% Of naïve T cells of all T cells (FACS analysis)	% Occupied by >0.001% clonotypes, of all reads
Mother A	F	36	13,656,054	12,513,543 (91.6)	2,044,290	1,708,037	2.7	55.0	19.4
Mother B	F	43	13,872,805	12,901,795 (93.0)	1,213,738	918,557	2.6	27.3	46.2
Mother C	F	43	11,167,059	10,038,463 (89.9)	2,180,886	1,978,745	3.0	39.4	38.2
Child A1	M	11	4,687,578	2,889,352 (61.6)	756,772	729,800	2.6	57.2	16.4
Child A2	M	9	4,202,419	2,467,388 (58.7)	558,173	535,283	2.4	43.0	28.2
Child B1	M	16	4,830,536	3,173,376 (65.7)	821,908	790,592	3.2	60.9	17.2
Child B2	M	10	4,615,093	3,009,961 (65.2)	545,730	517,410	2.3	40.1	34.3
Child C1	F	6	6,081,365	3,940,367 (64.8)	1,104,982	1,060,854	2.3	73.7	13.9
Child C2	M	16	4,564,646	3,008,748 (65.9)	785,164	760,572	2.6	63.7	19.7



Clone type	Clone Size
Rare	Less than 0.001%
Small	From 0.001% to 0.01%
Medium	From 0.01% to 0.1%
Large	From 0.1% to 1%
Hyperexpanded	More than 1%

FIGURE 1 | Representation of T cell clones of different size in individual TCR beta repertoires. Colored bars represent the share of clonal space occupied by clones of given type (classified by size) for each of the nine donors. Light green bars represent the share of naïve $CD27^{high}CD45RA^{high}$ T cells as determined by FACS analysis. Dashed lines indicate the share of low-frequency TCR beta clonotypes equivalent to this population in each individual, which were included in "low-frequency in-frame clonotypes" analysis. Square brackets indicate the share occupied by high-frequency T cell clones each representing >0.001% of all T cells.

that we used above for the approximate identification of the subset of naïve TCR beta clonotypes, we hypothesized that the most abundant clonotypes predominantly represent antigen-experienced T cell clones. We defined this population as clones representing >0.001% of all CDR3 sequences. Thus, the lower bound for this group was approximately an order of magnitude greater than the upper border set for the low-frequency clones in a given donor's T cell pool (**Figure 1**). Such delineation with a gap between the two subsets minimized "contamination" by naïve TCR beta clonotypes. Still, the pool of high-frequency in-frame clonotypes could contain a portion of naïve clonotypes with TCR beta CDR3 sequence variants of low complexity, that are repetitively produced in thymus due to the convergent recombination events and thus may be highly represented (15).

This set of the 2,803–8,285 most abundant clonotypes per individual cumulatively occupied 13.9–46.2% of the homeostatic T cell space in each donor. These high-frequency TCR beta clonotypes were generally characterized by increased variability in TRBV gene segment usage, and related and unrelated mother-child pairs were nearly indistinguishable (**Figure 2C**; **Figures 3A,B**, bars 5, 6).

OVERLAP OF TCR BETA REPERTOIRES FOR RELATED AND UNRELATED MOTHER-CHILD PAIRS

Several studies in recent years have revealed that unrelated individuals widely share TCR beta repertoires (13–15, 30–32). However, it is presently unclear whether the repertoires of haploidentical individuals are characterized by a higher level of overlap compared to unrelated donors. Additionally, for related mother-child pairs, shared TCR beta variants could conceal microchimeric T cell clones that have been physically shared across the placenta (see below).

To address these questions, we performed comparative analysis of TCR beta repertoire overlap for related and unrelated mother-child pairs by quantifying CDR3 variant identity at the amino acid level, at the nucleotide level, and at the nucleotide level in conjunction with identical TRBV and TRBJ gene segment usage (i.e., fully identical TCR beta chains). We measured overlaps separately for low-frequency and high-frequency in-frame clonotypes (as delineated in **Figure 1**), and all in-frame clonotypes. **Table 2** shows raw, non-normalized numbers of CDR3 variants shared on average by related and unrelated mother-child pairs.

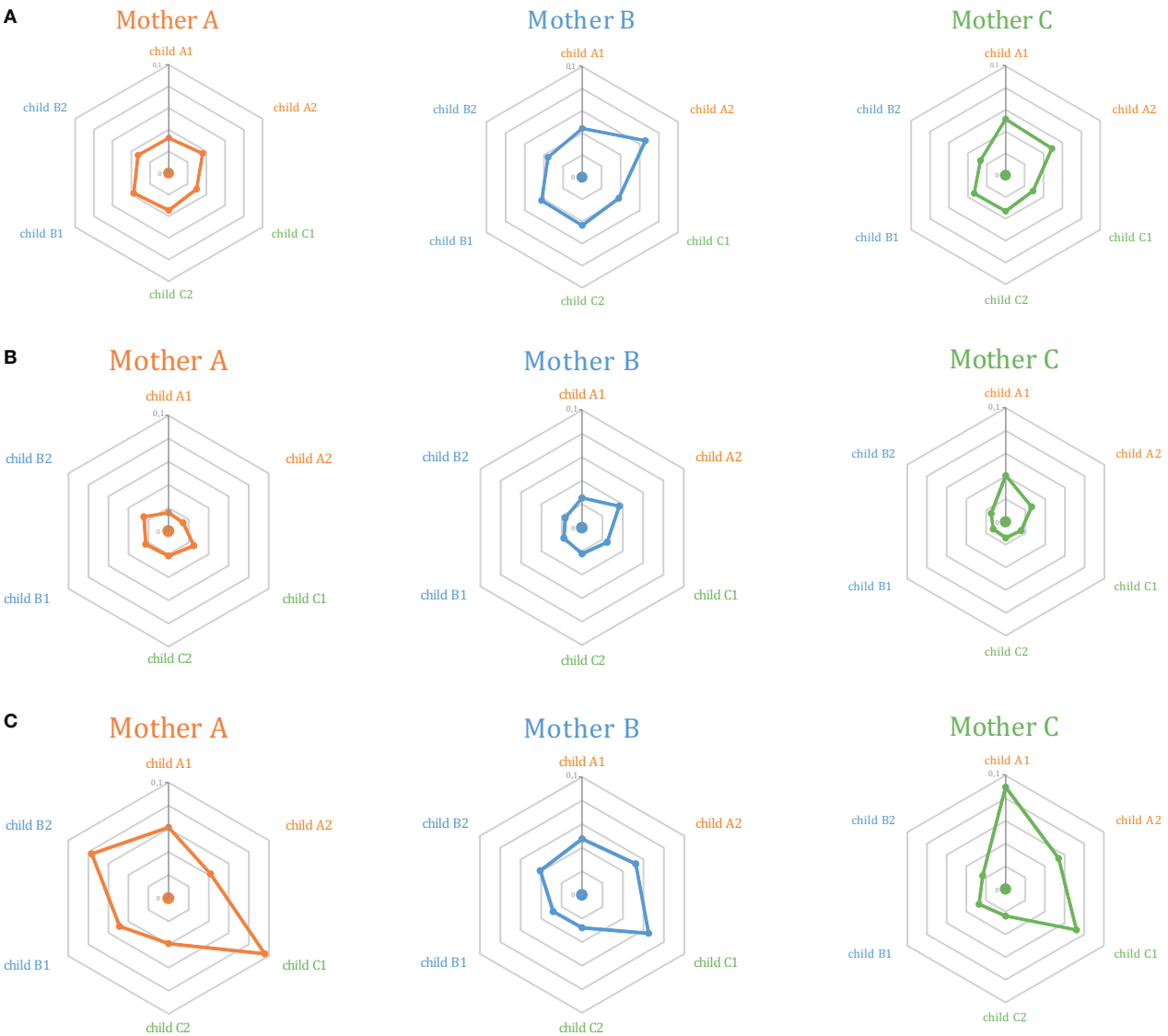


FIGURE 2 | Relative similarity of TRBV gene segments usage analyzed using the Jensen–Shannon divergence method for (A) out-of-frame TCR beta variants; (B) Low-frequency in-frame TCR beta clonotypes; and (C) high-frequency in-frame TCR beta clonotypes. The central dot in each diagram represents Mother **(A)** (orange), **(B)** (blue), or **(C)** (green). Surrounding dots represent the six children. Related children are shown in the same color as their mothers. The closer the "child" dot is to the central "mother" dot, the lower the Jensen–Shannon divergence (i.e., more similar TRBV gene segment usage).

For comparative analysis of relative overlap between subsets of different size, we normalized the number of identical CDR3 variants based on the sizes of the cross-compared samples as follows:

Formula 1: [normalized overlap between TCR sets A and B] = [overlap between TCR sets A and B]/([number of clonotypes in set A] × [number of clonotypes in set B]).

Normalized results are plotted in **Figure 5**. For all CDR3 categories, the degree of overlap was always slightly higher for related pairs, but this difference never approached a significant level compared to unrelated pairs. The highest level of overlap was observed for high-frequency clonotypes, in agreement with the previous work (15).

WITHIN AMINO ACID CDR3 OVERLAPS OF EXPANDED CLONOTYPES, PERCENTAGE OF CLONOTYPES WITH IDENTICAL TRBV GENES IS INCREASED FOR RELATED MOTHER-CHILD PAIRS

The CDR3 region is considered to form interactions mainly with antigenic peptide, while CDR1 and CDR2 encoded in the TRBV segment are mostly responsible for MHC recognition (33–35).

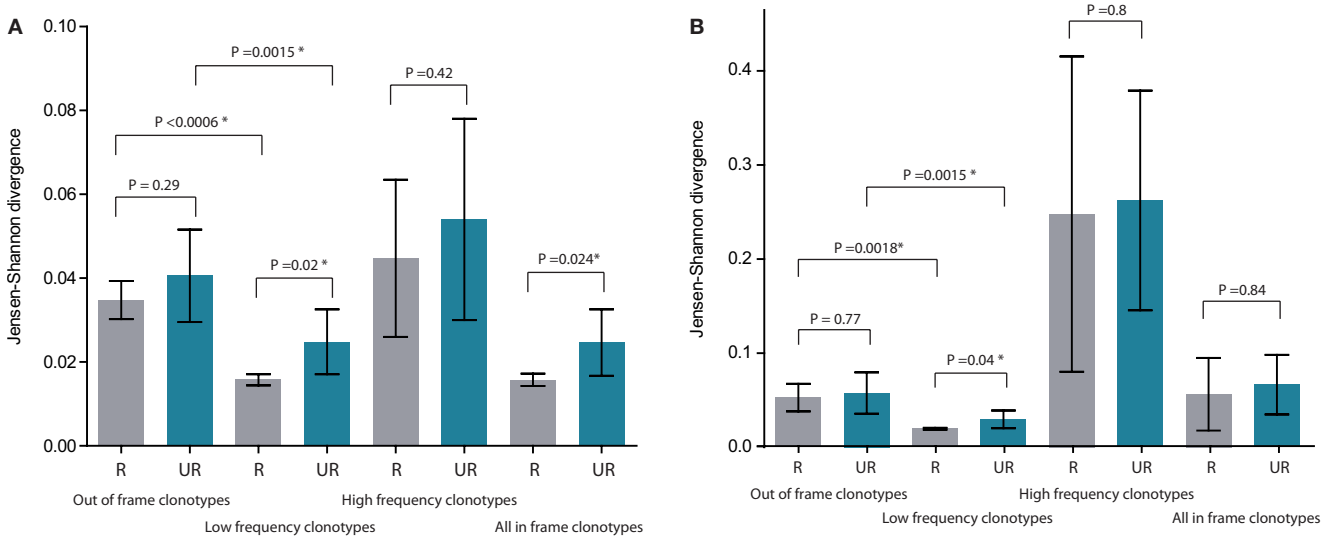


FIGURE 3 | Mean Jensen–Shannon divergence for TRBV gene segment usage for related (R, gray) versus unrelated (UR, blue) mother-child pairs, with SD. Bonferroni-corrected *P*-values are provided to account for multiple testing. **(A)** Comparison of TRBV gene segment usage frequencies at the level of distinct TCR beta clonotypes (but not sequencing reads, so that the influence of relative TCR beta clonotype size within a given repertoire is excluded). **(B)** Comparison of TRBV gene segment usage frequencies at the level of sequencing reads, i.e., considering relative abundance of clonotypes in peripheral blood. The lower the Jensen–Shannon divergence, the more similar the TRBV gene segment usage. *Denotes statistical significance. See **Figure 1** for delineation of low- and high-frequency clonotypes. Please note that "All in-frame clonotypes" include not only low-frequency and high-frequency clonotypes, but also the medium-frequency ones.

Some TRBV segments have nearly identical sequences taking part in CDR3 formation, so two different TRBV segments can often give rise to the same CDR3 amino acid sequence. However, in two individuals with similar or identical HLA alleles, proliferating antigen-specific clones with the same TRBV segment and CDR3 amino acid sequence that recognize the same peptide-MHC complex can be preferentially activated (36). Therefore, since related mother and child pairs share at least 50% of their HLA alleles, we could expect that antigen-experienced clones with identical amino acid CDR3 variants that recognize the same antigenic peptide should more often carry the same TRBV segment encoding CDR1 and CDR2 responsible for MHC recognition.

To verify this hypothesis, we analyzed various repertoire pairs comprising the 10,000 most abundant amino acid CDR3 clonotypes from each individual and computed overlap in terms of shared amino acid CDR3 sequences and shared amino acid CDR3 sequences carrying the same TRBV segment (i.e., identical CDR1, 2, and 3). We then determined the ratio of TRBV-CDR3 overlap to CDR3 overlap for each mother-child pair. In all cases, the ratio was greater for related mother-child pairs (1.3-fold, ± 0.16, **Figure 6A**). Moreover, we observed significant positive correlation of this ratio with the number of shared MHC-I alleles between individuals ($R = 0.62$, $P < 0.006$, **Figure 6B**; **Table 3**).

SELECTION IN THE THYMUS DECREASES AVERAGE CDR3 LENGTH COMPARED TO THE INITIALLY GENERATED REPERTOIRE

Comparison of the out-of-frame and in-frame CDR3 repertoires revealed that the former are characterized by higher average length (45.6 ± 0.4 versus 43.3 ± 0.2) and an increased number of added nucleotides (8.6 ± 0.2 versus 7.4 ± 0.1, see **Figure 7A**), in both mothers and children.

This finding indicates that, upon recombination, the initially generated TCR beta CDR3 repertoire (the parameters of which are preserved in the non-functional out-of-frame repertoire) is characterized by higher average length, while further selection in thymus essentially shapes the repertoire toward lower CDR3 length and fewer added nucleotides.

SEARCHING FOR MICROCHIMERIC CLONES TRANSFERRED ACROSS THE PLACENTA AS MATURE T CELLS

It is well established that mother and child exchange cells across the placenta during pregnancy (37–42), and that the progeny of these migrating cells persist in the new host for decades after gestation (43–45).

Most authors agree that lymphoid progenitor cells commonly cross the placenta to populate the new host (45–48). Some observations also indicate that mature T cells can transmigrate through the placenta (see Discussion). However, it remains to be determined whether the transferred mature T cells (hereinafter referred to as mature-microchimeric T cells) can further persist and serve as functional T cell clones in their new host.

We hypothesized that the present deep sequence analysis of such a substantial portion of the maternal and fetal TCR repertoire (including the absolute majority of proliferated antigen-experienced T cell clones) could reveal the presence of transferred and multiplied functional T cell populations, albeit without the immediate ability to distinguish the direction of transfer (i.e., maternal versus fetal microchimerism). Indeed,

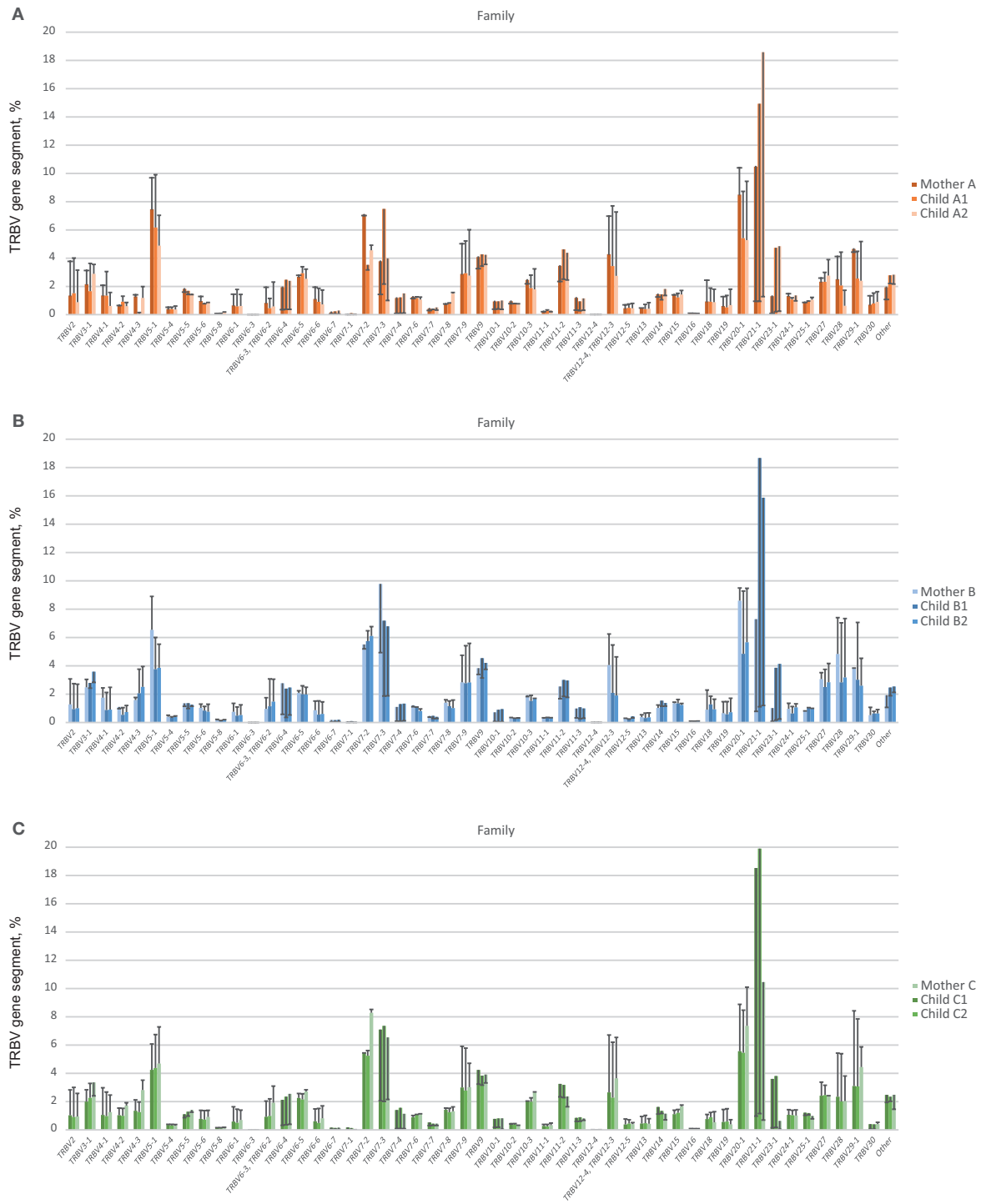


FIGURE 4 | TRBV gene segment usage in functional low-frequency TCR beta clonotypes in comparison to out-of-frame TCR variants. Colored bars indicate the representation of a particular TRBV gene segment family in out-of-frame TCR variants from each individual. Lines represent alterations in TRBV gene segment representation in functional low-frequency TCR beta clonotypes relative to out-of-frame TCR variants. **(A)**, **(B)**, and **(C)** depict TRBV gene segment usage for related donors from family **(A)**, **(B)**, and **(C)**, respectively.

microchimeric T cell clones that were initially transferred across the placenta as mature T cells (mature-microchimeric T cell clones) within a given mother-child pair should be characterized by the same TCR beta CDR3 nucleotide sequence and the same TRBV and TRBJ gene segments, which therefore could serve as a clone-specific identifier.

Table 2 | Average number of shared TCR beta CDR3 clonotypes in related and unrelated pairs.

Pairs	Amino acid			Nucleotide			Nucleotide TRBV, TRBJ identical		
	Low-frequency	High-frequency	All clonotypes	Low-frequency	High-frequency	All clonotypes	Low-frequency	High-frequency	All clonotypes
Related	71,714 ± 25,890	117 ± 34	83,257 ± 26,056	4,938 ± 2,237	69 ± 7	8,407 ± 3,396	2,015 ± 893	40 ± 10	3,456 ± 1,359
Unrelated	68,979 ± 18,822	99 ± 15	80,304 ± 18,626	4,640 ± 1,587	59 ± 12	7,955 ± 2,227	1,823 ± 624	31 ± 11	3,135 ± 897

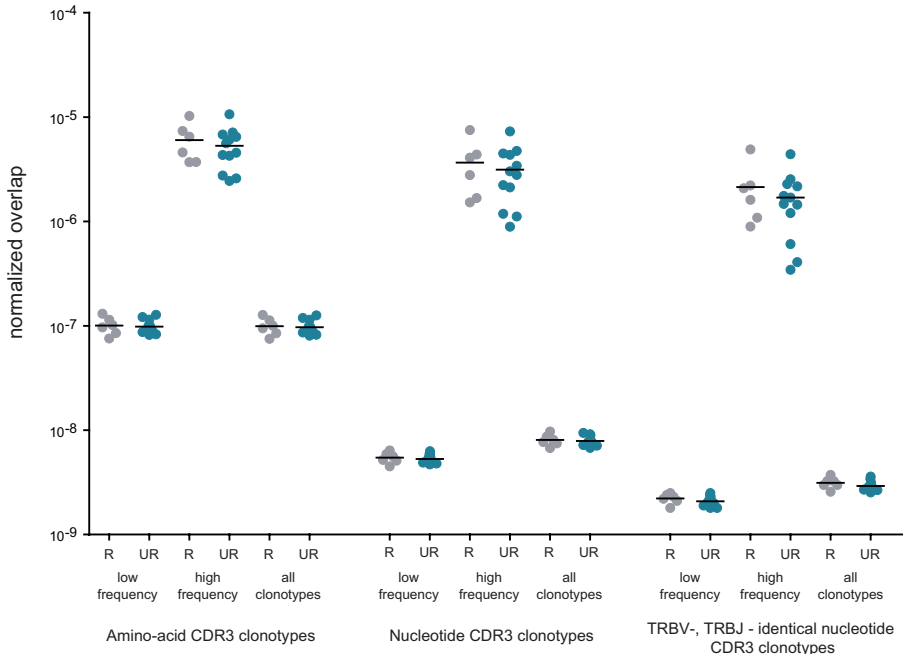


FIGURE 5 | Normalized overlap of individual TCR beta repertoires. Overlaps are shown at the level of CDR3 amino acid sequences, nucleotide sequences, and nucleotide sequences with identical TRBV and TRBJ segments; for related (gray) versus unrelated (blue) mother-child pairs; and for low-frequency, high-frequency, and all in-frame clonotypes. The number of intersections was normalized as described in Formula 1. R, related pairs; UR, unrelated pairs.

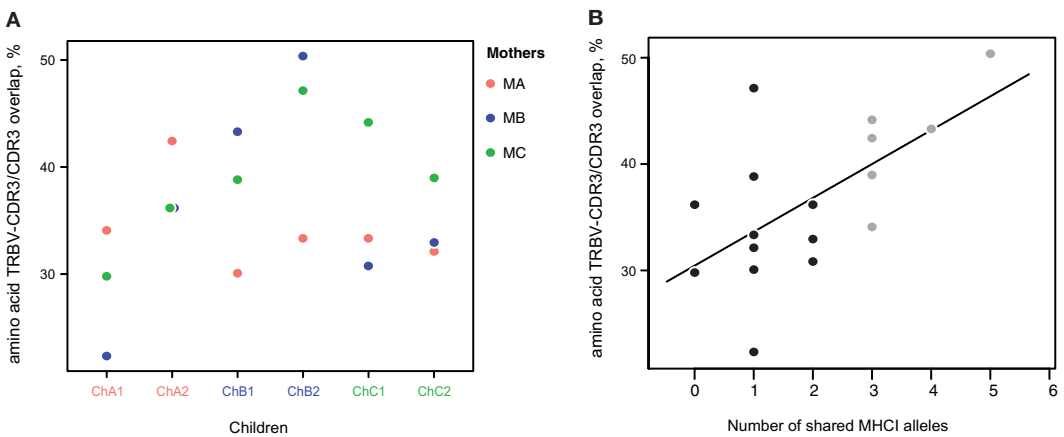


FIGURE 6 | Amino acid TRBV-CDR3/CDR3 overlap ratio. **(A)** The ratio of TRBV-CDR3 overlap to CDR3 overlap for all possible mother-child pairs, based on the 10,000 most highly represented clonotypes from each donor. Related mother-child pairs had a higher ratio relative to children with either of the unrelated mothers. **(B)** The number of shared MHC-I alleles in mother-child pairs correlates with the TRBV-CDR3/CDR3 overlap ratio for the 10,000 most abundant CDR3 clonotypes. Solid line displays linear regression fit; the Pearson correlation coefficient was 0.62 ($P < 0.006$). Related and unrelated pairs are shown in gray and black, respectively.

Table 3 | HLA typing.

	HLA-A	HLA-B	HLA-C	DRB1	DRB3	DRB4	DQB1
Mother A	02, 24	15, 57	03, 06	07, 14	01–03	01	03, 05
Mother B	02, 23	44, 51	02, 04	07, 11	01–03	01-02	02, 03
Mother C	01, 11	08, 35	04/08, 07	01, 03	01–03	–	02, 05
Child A1	02, 25	15, 18	03, 12	04, 14	01–03	01–02	03, 05
Child A2	02, 02	15, 44	03, 05	11, 14	01–03	–	03, 05
Child B1	02, 23	38, 44	04, 12	07, 13	01–03	01–02	02, 06
Child B2	02, 23	27, 44	02, 04	07, 11	01–03	01–02	02, 03
Child C1	02, 11	35, 38	04, 12	01, 13	01–03	–	05, 06
Child C2	02, 11	35, 38	04, 12	01, 13	01–03	-	05, 06

However, ~40% of the CDR3 nucleotide variants shared between any two individuals were characterized with the same TRBV and TRBJ gene segments, in similar numbers for both related and unrelated mother-child pairs. This means 1,766–5,410 shared clonotype variants across different donor pairs (**Table 2**; **Figure 5**). This widespread sharing of identical TCR beta nucleotide variants makes the TRBV-CDR3-TRBJ identifier insufficient to distinguish clones that were physically transferred across the placenta as mature T cells with recombined TCRs from public TCRs resulting from independent convergent recombination events (15, 32). Thus, if mature-microchimeric T cell clones are present, they are concealed amongst the overwhelming majority of natural public TCRs, and additional characteristics are needed to delineate them.

It has been reported that public TCR beta clonotypes are generally characterized by a low number of added nucleotides in CDR3 (i.e., low complexity) (14, 15, 32). We therefore used the number of added nucleotides as an additional selective characteristic that essentially determines the probability of convergent recombination events leading to CDR3 variants that are identical at the nucleotide level (32, 49). Comparison of this characteristic for all TCR beta CDR3 nucleotide variants and those TRBV-CDR3-TRBJ nucleotide variants that were shared between unrelated mother-child pairs revealed that the latter were characterized by much lower numbers of added nucleotides (**Figure 7A**).

The transfer of mature T cells across the placenta should not be dependent on CDR3 length or the number of added nucleotides. In humans, it has been demonstrated that there is no significant difference between adult blood and cord blood samples in the mean number of added nucleotides (50). Therefore, this characteristic should be essentially identical for both feto-maternal and materno-fetal mature-microchimeric T cell clones and for the general TCR beta repertoire. If the TCR beta repertoires of related mother-child pairs carry mature-microchimeric T cell clones of interest, we would expect to observe shaping of the added nucleotide curve proportional to the contribution of such clones to the repertoire overlap (**Figure 7B**).

The sensitivity of this method to the percentage of mature-microchimeric T cell clones in the shared TCR beta population is therefore limited by the natural dispersion of the added nucleotide curves for unrelated pairs. For example, if mature-microchimeric T cell clones contribute ~0.3% of the TRBV-CDR3-TRBJ overlap for a mother-child pair (i.e., ~10 out of 3,000 overlapping clonotypes, out of the $\sim 1 \times 10^6$ total clonotypes sequenced from each donor), the shape of the added nucleotide curve would be indistinguishable from that of an unrelated donor pair – and therefore below the sensitivity threshold of this method. In contrast, the presence of 100 mature-microchimeric T cell clones out of 3,000 clonotypes (i.e., 3.3% of shared variants) per pair of related donors could be clearly distinguished (**Figure 7B**), and this can therefore be considered as the approximate sensitivity limit of the method. We subsequently determined that the presence of mature-microchimeric T cell clones is undetectable in all cases, based on the added nucleotide curves for overlapping TRBV-CDR3-TRBJ nucleotide sequences for our six related mother-child pairs (**Figure 7C**). Correspondingly, the average numbers of added nucleotides in the shared TRBV-CDR3-TRBJ nucleotide variants were indistinguishable for related versus unrelated mother-child pairs (data not shown).

The above-described comparison of added nucleotide curves was performed at the level of distinct TCR beta clonotypes, but not sequencing reads, so that the influence of each T cell clone's relative representation within the repertoire was excluded. Similar albeit noisier results we have obtained when performing the same analysis at the level of sequencing reads (i.e., taking into account relative clonal size).

As such, we have not identified any meaningful difference between the subsets of shared TRBV-CDR3-TRBJ nucleotide variants for related versus unrelated mother-child pairs that would allow us to establish detection of a subpopulation of mature-microchimeric T cells that have been systemically shared during pregnancy as mature naïve or memory T cells, and which subsequently have engrafted and survived for years.

DISCUSSION

TRBV GENE USAGE

For out-of-frame TCR beta variants, which are not expressed and thus avoid any selection, TRBV gene usage was slightly more similar but generally comparable for related versus unrelated mother-child pairs (**Figures 2A** and **3**). This indicates that inherited maternal factors associated with the TCR recombination machinery are insufficient to yield the essentially similar TRBV gene segment selection in the child.

Remarkably, both within related and unrelated pairs, TRBV gene segment usage in low-frequency in-frame TCR beta clonotypes was more similar compared to that in the out-of-frame TCR beta variants (**Figure 3**). The equalization of the usage of TRBV gene segments in functional TCR variants (**Figure 4**) is probably a manifestation of selective pressure during thymic T cell selection, which should distinguish TRBV gene usage in functional TCRs from that preserved in unselected, out-of-frame TCR beta variants. This pressure on relative TRBV usage frequencies was prominent and led to significant convergence in both related ($P = 0.0006$) and unrelated ($P = 0.0015$) pairs, indicating that thymic selection essentially and similarly shapes the initial output of the TCR recombination machinery at the population level.

Interestingly, thymic selection also essentially filters out the longest CDR3 variants with large numbers of added nucleotides, as can be concluded from our comparison of non-functional out-of-frame and in-frame TCR beta CDR3 repertoires (**Figure 7A**).

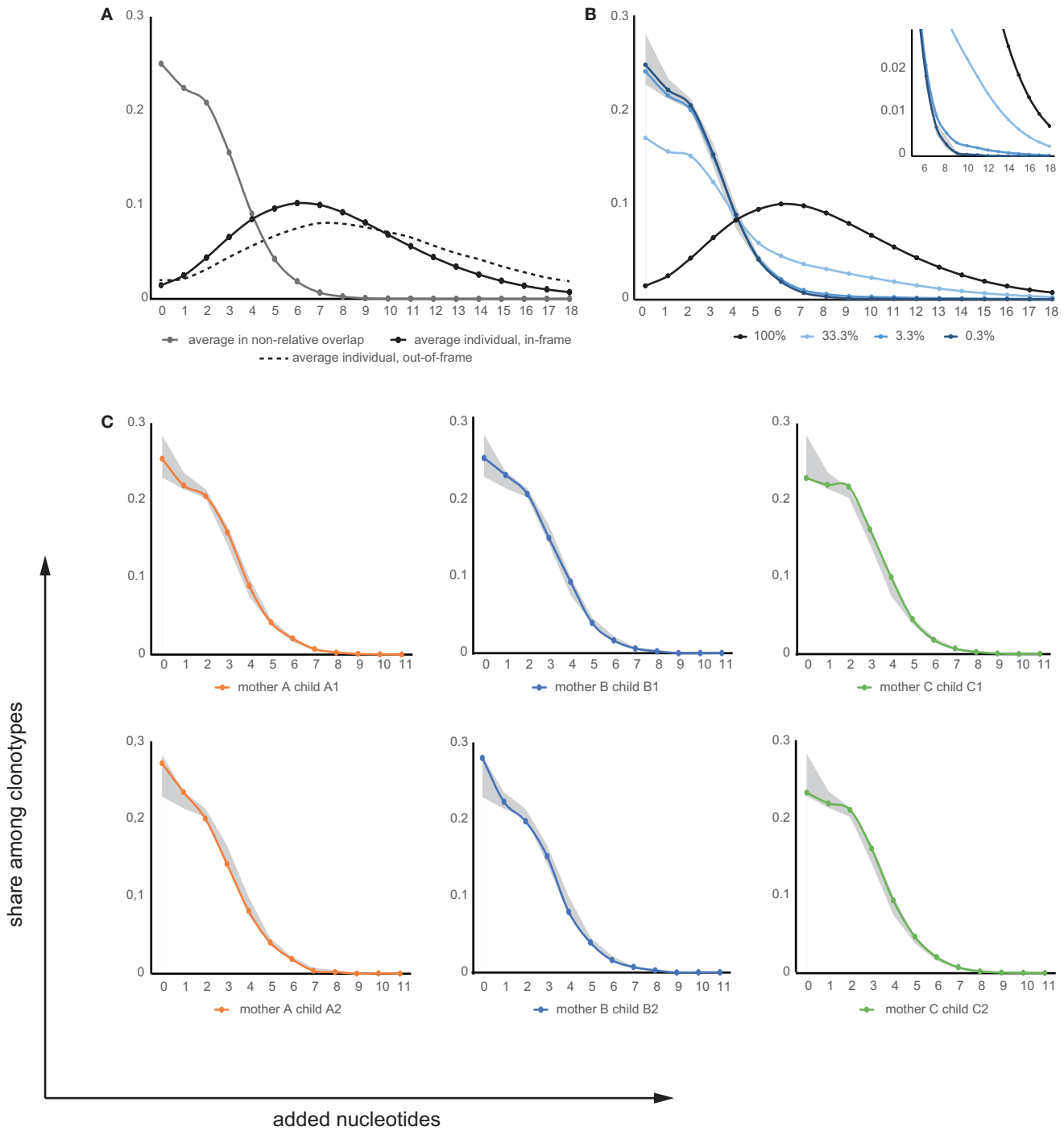


FIGURE 7 | Added nucleotide curves. **(A)** Average distribution of added nucleotides within CDR3 for individual TCR beta repertoires (in-frame: black solid line; out-of-frame: black dashed line) and for TCR beta clonotypes shared between unrelated individuals (gray). **(B)** Modeling of added nucleotide curves for shared TRBV-CDR3-TRBJ variants between mother-child pairs based on input of mature-microchimeric TCR beta CDR3 variants in different proportions. This was derived from the curves in **(A)**, which depict added nucleotide distributions for shared clonotypes between unrelated individuals (gray; equivalent to near-zero contribution to shared clonotypes) and for any individual repertoire (black; equivalent to 100% contribution to shared clonotypes), mixed in different proportions. Lines represent model input where mature-microchimeric TCR beta is equal to 100, 33, 3.3, or 0.3% of shared clonotypes. Shaded region shows the range for unrelated pairs. Inset shows a magnified view. **(C)** Added nucleotide curves for TRBV-CDR3-TRBJ variants shared in each related mother-child pair. Shaded region shows the range for unrelated pairs.

Since TRBV gene segments encode the fragments of TCR chains that interact with MHC (33–35), we would expect that related mother-child pairs, being haploidentical (i.e., sharing at least 50% of HLA alleles), are characterized by more similar TRBV gene segment usage frequencies at the level of functional T cells compared to unrelated donor pairs due to the impact of identical HLA genes in thymic selection. Indeed, we observed that differences in TRBV gene segment usage in related versus unrelated pairs became more pronounced and statistically significant ($P = 0.02$) at the level of low-frequency, in-frame TCR beta CDR3 clonotypes (**Figure 3**). However, the general direction of TCR beta repertoire shaping was similar for related and unrelated donors, suggesting that the pressure of thymic selection is relatively homogenous in the population. The strength of this general pressure was far greater relative to the specific changes that were characteristic of related donors, which only added a minor codirectional trend (**Figures 3** and **4**).

The subset of high-frequency TCR beta clonotypes was characterized by increased variability in TRBV segment usage, and related and unrelated mother-child pairs were indistinguishable at this level (**Figures 2C** and **3**). This is presumably due to the fact that different antigen specificities (but not TRBV segment interaction with MHC) play a dominant role in the priming and expansion of T cell clones, and this semi-random process negates the initial correlations that we observed in TRBV gene usage at the level of naïve T cells.

It should be noted, however, that the above analysis refers to low- and high-frequency clonotypes, which do not fully coincide with the naïve and antigen-experienced T cell subsets, respectively. It was previously demonstrated in other studies that recombinatorial biases might result in relatively high frequencies for certain naïve T cell clones, whereas some memory T cell clones may occur at relatively low frequencies (11, 14, 15). Moreover, these studies have shown a substantial overlap between the naïve and memory T cell repertoires, which suggests that a number of TCR beta CDR3 clonotypes could be associated with both subsets, being paired with either the same or alternative TCR alpha chains.

OVERLAP OF TCR BETA REPERTOIRES

We observed the greatest relative overlap of TCR beta repertoires among high-frequency clonotypes. This observation can be explained by the presence of common expanded antigen-experienced clonotypes recognizing the same antigens, as well as of high-frequency naïve clonotypes carrying the TCR beta CDR3 sequence variants of low complexity that are repetitively produced in thymus and may be highly represented both within and between individuals (15).

In all comparisons, only slightly higher numbers of shared clonotypes were observed in related versus unrelated mother-child pairs (**Figure 5**). This observation is in agreement with the previous report by Robins et al. where the overlap in the naïve CD8+ CDR3 sequence repertoires was suggested to be independent of the degree of HLA matching based on results obtained from three related donors (14). Here, we have achieved a more accurate comparison by studying a larger cohort of related donors, using unbiased library preparation techniques, sequencing the samples being compared on separate Illumina lanes to protect from potential cross-sample contamination on the solid phase and performing deeper individual profiling. Even with these various methodological improvements, we still observed only a subtle trend toward increased TCR beta repertoire overlap in related individuals.

However, among the shared high-frequency amino acid CDR3 variants, the percentage of TRBV-CDR3 identical clonotypes was always higher for related pairs compared to unrelated ones, and correlated with the number of identical MHC-I alleles (**Figure 6**). This finding indicates that optimal recognition of the particular peptide-MHC complex often requires full functional convergence of the TCR beta chain, leading to an increased share of TRBV-identical common CDR3 variants in individuals carrying the same HLA alleles. Notably, this phenomenon was observed for bulk T cell populations, where the input of CD8+ T cells was sufficient to provide correlation. This correlation would probably be much higher if we were to specifically analyze sorted CD8+ T cells.

SEARCHING FOR PERSISTENT MATURE-MICROCHIMERIC CLONES

In humans, maternal T cells are present in different fetal tissues (46, 48, 51), and may be present in the cord blood at a frequency of 0.1–0.5% of total T cells (48, 52). This can represent hundreds of thousands or millions of cells, of which many are likely to be memory T cells (52) capable of further clonal proliferation. Transmigration of maternal differentiated effector/memory Th1 and Th17 cells through the placenta was recently demonstrated in mouse models (53). Transfer of mature T cells is also possible in the opposite direction, and the presence of fetal microchimeric CD4+ and CD8+ T cells was registered in maternal blood during normal pregnancy in humans, predominantly in the third trimester (41) when mature α/β T cells are circulating in the fetus at significant numbers (54). Such mature-microchimeric T cell clones could further affect immunity to solid tumors (55, 56), influence transplantation tolerance (7), cause autoimmune diseases (3, 4, 43, 56–59), or protect the child against infections he/she has never encountered before.

Recent work has demonstrated that, in general, experienced clonal T cells commonly persist in the body for many years (17, 60). We have observed more than 20,000 TCR beta clonotypes that persisted in a patient for at least 7 years – from 2005 until 2012 – even after the patient underwent autologous HSC transplantation in 2009 [Ref. (16) and our unpublished data]. Similarly, naïve T cell clones persist in the body for many years after loss of thymus functionality (61). Therefore, if the engraftment of mature T cell clones transferred from mother to child and/or *vice versa* is a systemic process, we could expect to be able to verify the presence of such clones by using characteristic TCR beta CDR3 variants as clonal identifiers.

In our repertoire analysis, we did not observe mature-microchimeric T cell clones at a level of methodological sensitivity of ~100 mature-microchimeric clones per 10^6 analyzed TCR beta clonotypes. Still, this does not preclude the existence of mature T cell-based maternal or fetal microchimerism at levels below the sensitivity achieved in the current study, in minor number of individuals, or in pathological conditions such as autoimmune disease.

It should be noted that deep TCR beta profiling methodology presently appears to be insufficiently sensitive for identifying

particular expanded mature-microchimeric T cell clones, due to the general abundance of common identical TCR beta clonotypes. The following combination of methods could offer a potential way forward: (1) deep TCR beta profiling suggesting the presence of a particular expanded mature-microchimeric T cell clone, preferably with many added nucleotides within CDR3; (2) cell sorting using a TRBV family specific antibody in order to enrich for the hypothetical microchimeric clone of interest; and (3) real-time PCR confirmation of increased microchimerism in the sorted sample.

We also believe that further development of NGS profiling methods – especially in combination with the use of live cell-based emulsion PCR to identify paired TCR alpha-beta chains (62), and to potentially identify TCR beta chains paired with specific HLA molecules serving as an internal marker of microchimeric clones – should greatly facilitate future studies of mature T cell microchimerism in health and disease.

ACKNOWLEDGMENTS

We are grateful to M. Eisenstein for the English editing. This work was supported by the Molecular and Cell Biology program RAS, Russian Foundation for Basic Research (12-04-33139, 13-04-01124, 12-04-00229, 13-04-00998), and European Regional Development Fund (CZ.1.05/1.1.00/02.0068).

REFERENCES

1. Zenclussen AC. Adaptive immune responses during pregnancy. *Am J Reprod Immunol* (2013) **69**:291–303. doi:10.1111/aji.12097
2. Nelson JL, Furst DE, Maloney S, Gooley T, Evans PC, Smith A, et al. Microchimerism and HLA-compatible relationships of pregnancy in scleroderma. *Lancet* (1998) **351**:559–62. doi:10.1016/S0140-6736(97)08357-8
3. Sarkar K, Miller FW. Possible roles and determinants of microchimerism in autoimmune and other disorders. *Autoimmun Rev* (2004) **3**:454–63. doi:10.1016/j.autrev.2004.06.004
4. Lepez T, Vandewoestyne M, Hussain S, Van Nieuwerburgh F, Poppe K, Velkeniers B, et al. Fetal microchimeric cells in blood of women with an autoimmune thyroid disease. *PLoS One* (2011) **6**:e29646. doi:10.1371/journal.pone.0029646
5. Stern M, Ruggeri L, Mancusi A, Bernardo ME, De Angelis C, Bucher C, et al. Survival after T cell-depleted haploidentical stem cell transplantation is improved using the mother as donor. *Blood* (2008) **112**:2990–5. doi:10.1182/blood-2008-01-135285
6. Burlingham WJ, Benichou G. Bidirectional alloreactivity: a proposed microchimerism-based solution to the NIMA paradox. *Chimerism* (2012) **3**:29–36. doi:10.4161/chim.21668
7. Jankowska-Gan E, Sheka A, Sollinger HW, Pirsch JD, Hofmann RM, Haynes LD, et al. Pretransplant immune regulation predicts allograft outcome: bidirectional regulation correlates with excellent renal transplant function in living-related donor-recipient pairs. *Transplantation* (2012) **93**:283–90. doi:10.1097/TP.0b013e31823e46a0
8. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) **114**:4099–107. doi:10.1182/blood-2009-04-217604
9. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, et al. Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol Med* (2011) **3**:201–7. doi:10.1002/emmm.201100129
10. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* (2011) **12**:106. doi:10.1186/1471-2164-12-106
11. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
12. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* (2012) **42**:3073–83. doi:10.1002/eji.201242517
13. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? *Nat Rev Immunol* (2008) **8**:231–8. doi:10.1038/nri2260
14. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* (2010) **2**:47ra64. doi:10.1126/scitranslmed.3001442
15. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, et al. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J Immunol* (2011) **186**:4285–94. doi:10.4049/jimmunol.1003898
16. Britanova OV, Bochkova AG, Staroverov DB, Fedorenko DA, Bolotin DA, Mamedov IZ, et al. First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. *Bone Marrow Transplant* (2012) **47**:1479–81. doi:10.1038/bmt.2012.44
17. Klarenbeek PL, Remmerswaal EB, ten Berge IJ, Doorenspleet ME, Van Schaik BD, Esveldt RE, et al. Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. *PLoS Pathog* (2012) **8**:e1002889. doi:10.1371/journal.ppat.1002889
18. Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, et al. Preparing unbiased T cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol* (2013) **4**:456. doi:10.3389/fimmu.2013.00456
19. Nemazee D. Receptor editing in lymphocyte development and central tolerance. *Nat Rev Immunol* (2006) **6**:728–40. doi:10.1038/nri1939
20. Venturi V, Rudd BD, Davenport MP. Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr Opin Immunol* (2013) **25**(5):639–45. doi:10.1016/j.coi.2013.07.001
21. Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL, et al. A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J Immunol* (2002) **168**:3099–104.
22. Britanova OV, Staroverov DB, Chkalina AV, Kotlobay AA, Zvezdova ES, Bochkova AG, et al. Single high-dose treatment with glucosaminyl-muramyl dipeptide is ineffective in treating ankylosing spondylitis. *Rheumatol Int* (2011) **31**:1101–3. doi:10.1007/s00296-010-1663-3
23. Lefranc MP, Giudicelli V, Busin C, Malik A, Mougenot I, Dehais P, et al. LIGM-DB/IMGT: an integrated database of Ig and TcR, part of the immunogenetics database. *Ann N Y Acad Sci* (1995) **764**:47–9. doi:10.1111/j.1749-6632.1995.tb55805.x
24. Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, et al. MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* (2013) **10**(9):813–4. doi:10.1038/nmeth.2555
25. Lin JT. Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* (1991) **37**:145–51. doi:10.1109/18.61115
26. Wang J, Vock VM, Li S, Olivas OR, Wilkinson MF. A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation. *J Biol Chem* (2002) **277**:18489–93. doi:10.1074/jbc.M111781200
27. Bhalla AD, Gudikote JP, Wang J, Chan WK, Chang YF, Olivas OR, et al. Nonsense codons trigger an RNA partitioning shift. *J Biol Chem* (2009) **284**:4062–72. doi:10.1074/jbc.M805193200
28. Favre D, Stoddart CA, Emu B, Hoh R, Martin JN, Hecht FM, et al. HIV disease progression correlates with the generation of dysfunctional naive CD8(low) T cells. *Blood* (2011) **117**:2189–99. doi:10.1182/blood-2010-06-288035
29. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. *Science* (1999) **286**:958–61. doi:10.1126/science.286.5441.958
30. Venturi V, Chin HY, Asher TE, Ladell K, Scheinberg P, Bornstein E, et al. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J Immunol* (2008) **181**:7853–62.
31. Li H, Ye C, Ji G, Han J. Determinants of public T cell responses. *Cell Res* (2012) **22**:33–42. doi:10.1038/cr.2012.1
32. Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y, et al. Recombinatorial biases and convergent recombination determine interindividual TCRbeta sharing in murine thymocytes. *J Immunol* (2012) **189**:2404–13. doi:10.4049/jimmunol.1102087

33. Hogquist KA, Baldwin TA, Jameson SC. Central tolerance: learning self-control in the thymus. *Nat Rev Immunol* (2005) **5**:772–82. doi:10.1038/nri1707
34. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* (2006) **24**:419–66. doi:10.1146/annurev.immunol.23.021704.115658
35. Garcia KC, Adams JJ, Feng D, Ely LK. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* (2009) **10**:143–7. doi:10.1038/ni.f.219
36. Miles JJ, Douek DC, Price DA. Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* (2011) **89**:375–87. doi:10.1038/icb.2010.139
37. Desai RG, Creger WP. Maternofetal passage of leukocytes and platelets in man. *Blood* (1963) **21**:665–73.
38. Herzenberg LA, Bianchi DW, Schroder J, Cann HM, Iverson GM. Fetal cells in the blood of pregnant women: detection and enrichment by fluorescence-activated cell sorting. *Proc Natl Acad Sci U S A* (1979) **76**:1453–5. doi:10.1073/pnas.76.3.1453
39. Iverson GM, Bianchi DW, Cann HM, Herzenberg LA. Detection and isolation of fetal cells from maternal blood using the flourescence-activated cell sorter (FACS). *Prenat Diagn* (1981) **1**:61–73. doi:10.1002/pd.1970010111
40. Nelson JL. Your cells are my cells. *Sci Am* (2008) **298**:64–71.
41. Adams Waldorf KM, Gammill HS, Lucas J, Aydelotte TM, Leisenring WM, Lambert NC, et al. Dynamic changes in fetal microchimerism in maternal peripheral blood mononuclear cells, CD4+ and CD8+ cells in normal pregnancy. *Placenta* (2010) **31**:589–94. doi:10.1016/j.placenta.2010.04.013
42. Nelson JL. The otherness of self: microchimerism in health and disease. *Trends Immunol* (2012) **33**:421–7. doi:10.1016/j.it.2012.03.002
43. Evans PC, Lambert N, Maloney S, Furst DE, Moore JM, Nelson JL. Long-term fetal microchimerism in peripheral blood mononuclear cell subsets in healthy women and women with scleroderma. *Blood* (1999) **93**:2033–7.
44. Maloney S, Smith A, Furst DE, Myerson D, Rupert K, Evans PC, et al. Microchimerism of maternal origin persists into adult life. *J Clin Invest* (1999) **104**:41–7. doi:10.1172/JCI6611
45. Loubiere LS, Lambert NC, Flinn LJ, Erickson TD, Yan Z, Guthrie KA, et al. Maternal microchimerism in healthy adults in lymphocytes, monocyte/macrophages and NK cells. *Lab Invest* (2006) **86**:1185–92.
46. Gotherstrom C, Johnsson AM, Mattsson J, Papadogiannakis N, Westgren M. Identification of maternal hematopoietic cells in a 2nd-trimester fetus. *Fetal Diagn Ther* (2005) **20**:355–8. doi:10.1159/000086812
47. Khosrotehrani K, Leduc M, Bachy V, Nguyen Huu S, Oster M, Abbas A, et al. Pregnancy allows the transfer and differentiation of fetal lymphoid progenitors into functional T and B cells in mothers. *J Immunol* (2008) **180**:889–97.
48. Mold JE, Michaelsson J, Burt TD, Muench MO, Beckerman KP, Busch MP, et al. Maternal alloantigens promote the development of tolerogenic fetal regulatory T cells in utero. *Science* (2008) **322**:1562–5. doi:10.1126/science.1164511
49. Murugan A, Mora T, Walczak AM, Callan CG Jr. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* (2012) **109**:16161–6. doi:10.1073/pnas.1212755109
50. Hall MA, Reid JL, Lanchbury JS. The distribution of human TCR junctional region lengths shifts with age in both CD4 and CD8 T cells. *Int Immunol* (1998) **10**:1407–19. doi:10.1093/intimm/10.10.1407
51. Jonsson AM, Uzunel M, Gotherstrom C, Papadogiannakis N, Westgren M. Maternal microchimerism in human fetal tissues. *Am J Obstet Gynecol* (2008) **198**(325):e321–6.
52. Burlingham WJ, Nelson JL. Microchimerism in cord blood: mother as anticancer drug. *Proc Natl Acad Sci U S A* (2012) **109**:2190–1. doi:10.1073/pnas.1120857109
53. Wienecke J, Hebel K, Hegel KJ, Pierau M, Brune T, Reinhold D, et al. Pro-inflammatory effector Th cells transmigrate through anti-inflammatory environments into the murine fetus. *Placenta* (2012) **33**:39–46. doi:10.1016/j.placenta.2011.10.014
54. Haynes BF, Martin ME, Kay HH, Kurtzberg J. Early events in human T cell ontogeny. Phenotypic characterization and immunohistologic localization of T cell precursors in early human fetal tissues. *J Exp Med* (1988) **168**:1061–80. doi:10.1084/jem.168.3.1061
55. Gadi VK. Fetal microchimerism in breast from women with and without breast cancer. *Breast Cancer Res Treat* (2010) **121**:241–4. doi:10.1007/s10549-009-0548-1
56. Fugazzola L, Cirello V, Beck-Peccoz P. Fetal microchimerism as an explanation of disease. *Nat Rev Endocrinol* (2011) **7**:89–97. doi:10.1038/nrendo.2010.216
57. Willer CJ, Sadovnick AD, Ebers GC. Microchimerism in autoimmunity and transplantation: potential relevance to multiple sclerosis. *J Neuroimmunol* (2002) **126**:126–33. doi:10.1016/S0165-5728(02)00048-6
58. Adams KM, Nelson JL. Microchimerism: an investigative frontier in autoimmunity and transplantation. *JAMA* (2004) **291**:1127–31. doi:10.1001/jama.291.9.1127
59. Lambert NC, Erickson TD, Yan Z, Pang JM, Guthrie KA, Furst DE, et al. Quantification of maternal microchimerism by HLA-specific real-time polymerase chain reaction: studies of healthy women and women with scleroderma. *Arthritis Rheum* (2004) **50**:906–14. doi:10.1002/art.20200
60. Naumova EN, Gorski J, Naumov YN. Two compensatory pathways maintain long-term stability and diversity in CD8 T cell memory repertoires. *J Immunol* (2009) **183**:2851–8. doi:10.4049/jimmunol.0900162
61. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mogling R, De Boer AB, et al. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* (2012) **36**:288–97. doi:10.1016/j.immuni.2012.02.006
62. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* (2013) **43**(9):2507–15. doi:10.1002/eji.201343453

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 July 2013; accepted: 03 December 2013; published online: 25 December 2013.

*Citation: Putintseva EV, Britanova OV, Staroverov DB, Merzlyak EM, Turchaninova MA, Shugay M, Bolotin DA, Pogorelyy MV, Mamedov IZ, Bobrynina V, Maschan M, Lebedev YB and Chudakov DM (2013) Mother and child T cell receptor repertoires: deep profiling study. Front. Immunol. **4**:463. doi: 10.3389/fimmu.2013.00463*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Putintseva, Britanova, Staroverov, Merzlyak, Turchaninova, Shugay, Bolotin, Pogorelyy, Mamedov, Bobrynina, Maschan, Lebedev and Chudakov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 11 December 2013
doi: 10.3389/fimmu.2013.00439

Mathematical models of the impact of IL2 modulation therapies on T cell dynamics

Kalet León*, Karina García-Martínez and Tania Carmenate

Systems Biology Department, Center of Molecular Immunology, Habana, Cuba

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Tomasz Zal, University of Texas MD Anderson Cancer Center, USA
Joseph Reynolds, University of Leeds, UK

****Correspondence:***
Kalet León, Systems Biology Department, Center of Molecular Immunology, 216 Street, PO Box 16040, Atabey, Habana 11600, Cuba
e-mail: kalet@cim.sld.cu

Several reports in the literature have drawn a complex picture of the effect of treatments aiming to modulate IL2 activity *in vivo*. They seem to promote either immunity or tolerance, probably depending on the specific context, dose, and timing of their application. Such complexity might derive from the pleiotropic role of IL2 in T cell dynamics. To theoretically address the latter possibility, our group has developed several mathematical models for Helper, Regulatory, and Memory T cell population dynamics, which account for most well-known facts concerning their relationship with IL2. We have simulated the effect of several types of therapies, including the injection of: IL2; antibodies anti-IL2; IL2/anti-IL2 immune-complexes; and mutant variants of IL2. We studied the qualitative and quantitative conditions of dose and timing for these treatments which allow them to potentiate either immunity or tolerance. Our results provide reasonable explanations for the existent pre-clinical and clinical data, predict some novel treatments, and further provide interesting practical guidelines to optimize the future application of these types of treatments.

Keywords: mathematical model, T cell dynamics, interleukin 2, interleukin 2 mutants, regulatory T cells

INTRODUCTION

Several reports in the literature have drawn a complex picture of the effect of treatments aiming to modulate IL2 activity *in vivo*. These treatments seem to promote either immunity or tolerance, probably depending on the specific context, dose, and timing of their application.

Treatments that increase IL2 activity, simply by injecting it, have been shown to potentiate the immune response to vaccines (1–4) and are a current medical practice to enhance the natural anti-tumor immunity in patients with melanoma. However, several reports in the literature have shown that HIV (5–8) and melanoma (9) patients treated with IL2, experience an increase in $CD4^+$ $CD25^+$ $FoxP3^+$ regulatory T cells, which typically mediate natural immune tolerance. Moreover, several pre-clinical studies have further documented a tolerogenic effect of IL2. Injections of IL2 have been shown to prevent or ameliorate autoimmune responses in mice (10–12). Treatments which reduce natural IL2 activity, by sequestering it with anti-IL2 monoclonal antibodies, have been shown to induce autoimmune responses (13). And treatments intending to block IL2 activity, with non-depleting anti-IL2-receptor antibodies, are showed to have anti-tumoral effects (14). Nevertheless, in the clinical practice non-depleting anti-IL2-receptor antibodies are used to ameliorate the autoimmune reaction in patients with neoplasia, autoimmune diseases, and organ allograft rejection (15).

Further complexity to the latter picture has been recently added with the pre-clinical assessment of treatments based on immune-complexes formed by IL2 and monoclonal antibodies anti-IL2. This treatment shows a much more potent *in vivo* effect than IL2 alone, appears again to potentiate either immunity (16, 17) or tolerance (18), depending on the specific antibody used to form the immune-complexes. In particular, the specific epitope in the IL2 recognized by the antibody has been postulated as critical for this phenomenon (19, 20).

IL2 interacts with many different cells types, which express the three known chains of the IL2 receptor. Particularly relevant and complex is its relationship with the population dynamics of the CD4 T lymphocytes. IL2 was originally described as a potent $CD4^+$ T cell growth factor (21), which should in consequence enhance overall T cell immunity. However, several experiments have shown lately a critical role for this cytokine in the survival and proliferation of the $CD4^+CD25^+FoxP3^+$ T cells (regulatory T cells) (22, 23), which mediate the maintenance of natural and induced tolerance. The $CD4^+CD25^-FoxP3^-$ T cells (helper T cells) have been identified as the principal source of IL2 *in vivo* (24), suggesting that the regulatory T cells have to sequester the IL2 produced by these cells in order to proliferate and survive (25). Moreover, *in vitro and in vivo* experiments have shown that regulatory T cells inhibit the production of IL2 by the helper T cells (26), limiting in this way their own source of this essential cytokine. Thus, overall, it seems that IL2 has a dual role on its circuit of interactions with $CD4^+$ T cells. It could promote the proliferation of the helper T cells, which may drive effective immunity and foster IL2 production. But, it could also promote the expansion of regulatory T cells, which may turn off the immune reaction, as well as the IL2 production on its own. The dynamic balance between these opposite forces might explain the complexity observed in the effect of treatments that modulate IL2 activity, either sequestering it or further increasing it.

To theoretically address the latter hypothesis, our group has developed mathematical models for Helper, Regulatory, and Memory T cells dynamics, which account for most well-known facts

relative to their relationship with IL2. We have simulated the effect of several types of therapies including the injection of: IL2; antibodies anti-IL2; IL2/anti-IL2 immune-complexes, and mutants variants of IL2. We studied the qualitative and quantitative conditions of dose and timing for these treatments which allow them to potentiate either immunity or tolerance. Our results provide reasonable explanations for the existent pre-clinical and clinical data, predict some novel treatments, and further provide interesting practical guidelines to optimize the future application of these types of treatments.

MATERIALS AND METHODS

INTRODUCTION TO THE MATHEMATICAL MODEL

The mathematical model used in this paper is based on the one developed in Ref. (27) to describe the interaction between IL2 and helper (E) and regulatory (R) $CD4^+$ T cells and memory $CD8^+$ T cells inside a lymph node. The model includes several physical compartments, which minimally capture the bio-distribution of T cells, IL2, and antibodies in the immune system (see **Figure 1**). It includes several compartments, which represent different lymph nodes, where T cells are confined interacting with each other's, with the antigen presenting cells (APCs) and available soluble molecules. It includes also a compartment representing the blood (i.e., the circulatory system), which contains only soluble molecules, IL2, mutant variants of IL2 or anti-IL2 antibodies. Each lymph node in the system is connected to the blood compartment, allowing the free exchange of these soluble molecules.

DYNAMICS IN THE BLOOD COMPARTMENT

The concentration of soluble molecules in the blood compartment is assumed to decay with a constant characteristic rate, which represent renal elimination in the kidney. An external source term for these molecules is added in this compartment to simulate particular treatment applications. Interaction between free IL2 and anti-IL2 antibodies are modeled in this and other compartments as a dynamic equilibrium characterized by a given biding affinity. Equations for the dynamics in this compartment are presented in "Dynamics in the Blood Compartment" in Appendix A.

DYNAMICS FOR T CELLS INSIDE LYMPH NODES

The model includes, inside the lymph nodes, the dynamics of Helper (E), and Regulatory (R) T cells on three different functional states of their life cycle: resting, activated, and cycling cells. All the interactions involving these T cells occur in the presence of a constant amount of their cognate APCs and relevant homeostatic cytokines. The basic processes and interactions included in the model dynamics for these T cells are (see **Figure 2** and (27, 28) for a more detailed biological explanation, including references to experiments that sustained their validity):

i. Resting E and R cells are produced at constant rate by the thymus; they die with a constant decay rate; they get activated (becoming an activated cell) following conjugation to their cognate APCs. The activation of E cells can be inhibited by the presence of co-localized R cells on the APCs.
ii. The activated E and R cells could become cycling cells following a dose-dependent response to cytokine derived signals. The activated R cells get this signal from the interaction with available IL2 while the E cells could additionally use other homeostatic cytokines[1], which are referred in the model as

[1]Note that, although other cytokines are able to stimulate Tregs *in vitro*, several reports in the literature have indicated IL2 as the key cytokine for the proliferation

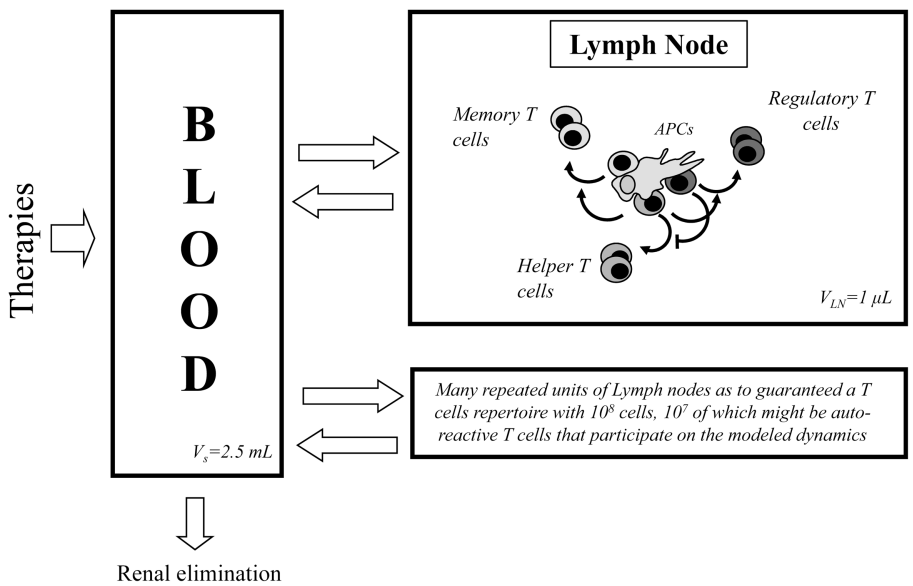


FIGURE 1 | Diagram of the processes occurring in the two compartments considered in the model. At the left side of the diagram the blood compartment is shown, where soluble molecules related with IL2 modulatory therapies are introduced and eliminated. This compartment is in constant molecular exchange with the lymph nodes (right side of the diagram). In this last class of compartment, occur the processes related with the dynamics of T cells and their interaction with the IL2 and other soluble molecules.

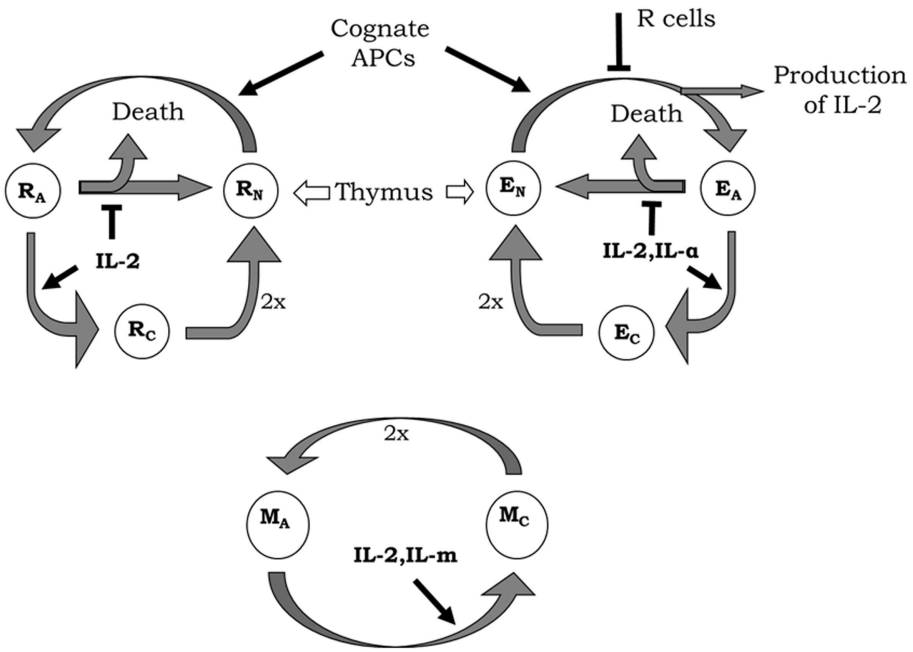


FIGURE 2 | Diagrams of helper (E), regulatory (R), and memory (M) T-cell life cycle considered in the model. New resting E (E_N) and R (R_N) cells are constantly generated by the thymus. These resting T cells become activated by interaction with their cognate APCs. During activation, E cells produce IL2, although the whole process can be inhibited by the presence of co-localized R cells. Activated E (E_A) and R (R_A) enter the cell cycle (becoming cycling cells) when receiving enough signal from IL2 or another external cytokine (IL-α) in the case of E cells. In the absence of enough cytokines, activated T cells become inactivated, where a fraction of cells simply returns back to the resting state and the other dies. Cycling E (E_C) and R (R_C) cells divide with a constant rate generating two new resting E or R cells, respectively. Memory T cells are assumed as being always in a sort of naturally activated state (even without any strong cognate interaction with APCs). Activated M (M_A) cells enter the cell cycle when receiving enough signals from IL2 or another external cytokine (IL-m). Cycling M cells (M_C) divide generating two new activated M cells.

ILa and are available inside the lymph node in a constant but limited amount. In the absence of enough cytokine derived signal, a fraction of the activated E or R cells revert to the resting state and the remaining fraction just die.

iii. The cycling E and R cells are fully committed to divide, producing two new resting cells. Thus, they are presumed to do so with a constant rate.

The model includes also the dynamics of a generic population of non-CD4 T cells, which binds weakly to the existent APCs, but proliferates in response to IL2 signal, with similar sensitivity than the activated helper $CD4^+$ T cells. This cells (referred as M cells) represent, the memory $CD8^+CD44^+$ T cells, which can proliferate in response to IL2 without any requirements of activation by cognate APCs (see **Figure 2**).

The dynamics of the number of T cells in the lymph node compartment, following the process described above, are modeled with the set of equations presented in "Dynamics of T Cells in the Lymph Node Compartment" in Appendix B.

DYNAMICS IL2 AND ANTIBODIES ANTI-IL2 INSIDE THE LYMPH NODE

The dynamics of IL2 molecules inside the lymph node takes into account the role of T cells in the production and degradation of this cytokine. The following processes are considered in the model [see **Figure 2** and Ref. (27) for a more detailed biological explanation, including references to experiments that sustained their validity]:

iv. IL2 is produced by E cells upon activation. It is produced as a burst whenever a resting E cell becomes an activated E cell. Such production of IL2 is inhibited, together with the E cell activation, by the presence of co-localized R cells on the APCs.

v. IL2 is degraded in the lymph nodes, after being internalized by the T cells in the form of complexes with the IL2 receptor at their cell surface.

Interactions of IL2 and T cells in the model are based on the expression by these cells, either in the resting, activated or cycling state, of different levels of the IL2 receptor. These receptors mediate the binding of IL2, which provide a stimulatory signal in a dose-dependent fashion to the T cell. In this model the three known chains of the IL2 receptor, alpha, beta, and gamma (29) are included. These three chains are

and survival of Treg cells *in vivo*. The group of Freitas (24) have shown that the absence of $CD4^+$ T cells capable of producing IL2, leads to the absence of Treg cells and to the development of autoimmunity. Moreover, mice knockouts of IL2 or IL2 receptor components have been shown to lack the accumulation of Tregs *in vivo*, exhibiting once more an autoimmune phenotype (48, 49). Interestingly in these latter scenarios of autoimmune mice, other cytokines besides IL2 are capable to maintain and expand the auto-reactive helper $CD4^+$ T cells (perhaps IL7, IL15, or IL21).

combined dynamically at the cell surface, upon IL2 binding, to conform the two known signaling forms of the IL2 receptors. The following processes and known facts are considered in the model regarding this interaction [see **Figure 3** and Ref. (27)]:

vi. IL2/IL2Receptor complexes formation is modeled as a multi-step process: free, soluble, IL2 binds initially to the available free alpha or free beta chains of the receptor, and only then can form dimers or trimers with the remaining IL2 receptor chains at the cell membrane. The gamma chain is assumed to be always in excess compared with the amount of beta chain bound to IL2, either alone or together with alpha chain. Therefore gamma chain joins immediately to these membrane complexes, forming the well known intermediate (beta-gamma-IL2) or high affinity (alpha-beta-gamma-IL2) IL2–IL2 receptor complexes.

vii. IL2/IL2Receptor configurations, which include the beta and gamma chains (high-affinity alpha-beta-gamma, and intermediate affinity beta-gamma receptor), trigger a signal into the T cells (19). Therefore, in the model, the mean number of such signaling receptors per activated E cell, R cell, and M cell are counted. Then, the probability of getting enough signal as to become a cycling cell, for any particular activated E, R, or M cell, is computed with a sigmoid dose response curve, of the mean signaling level. The use of a sigmoid dose response curve is based on direct experimental observations on *in vitro* culture of $CD4^+$ T cells (30) stimulated with recombinant IL2.

viii. Beta and gamma chain of the IL2 receptor are similarly expressed by E and R cells in all functional states, but the expression of the alpha chain is modulated with T cell activation (31). R cells constitutively express the alpha chain in the resting state, but further increase its expression level with activation. E cells do not express the alpha chain in the resting state, but gain a significant expression level with activation.

ix. The M cells are assumed to express a negligible amount of the alpha chain of IL2 receptor, but have levels of the beta and gamma chain which are higher than those of helper and regulatory T cells (32).

Antibodies anti-IL2 are modeled as molecules that can form complexes with the IL2, blocking or not its binding to the different chains of the IL2 receptor at the T cell surface. IL-2 mutants are modeled as a molecule bearing similar properties than wild-type IL-2, but differing in some specific parameter value on each case. In particular, we simulate the effects of IL2 mutants with an either reduced or increased Kon for the alpha or beta chains of the IL2R.

The equations in the model describing the dynamics of the number of molecules circulating in the Lymph Node (IL2, anti-IL2

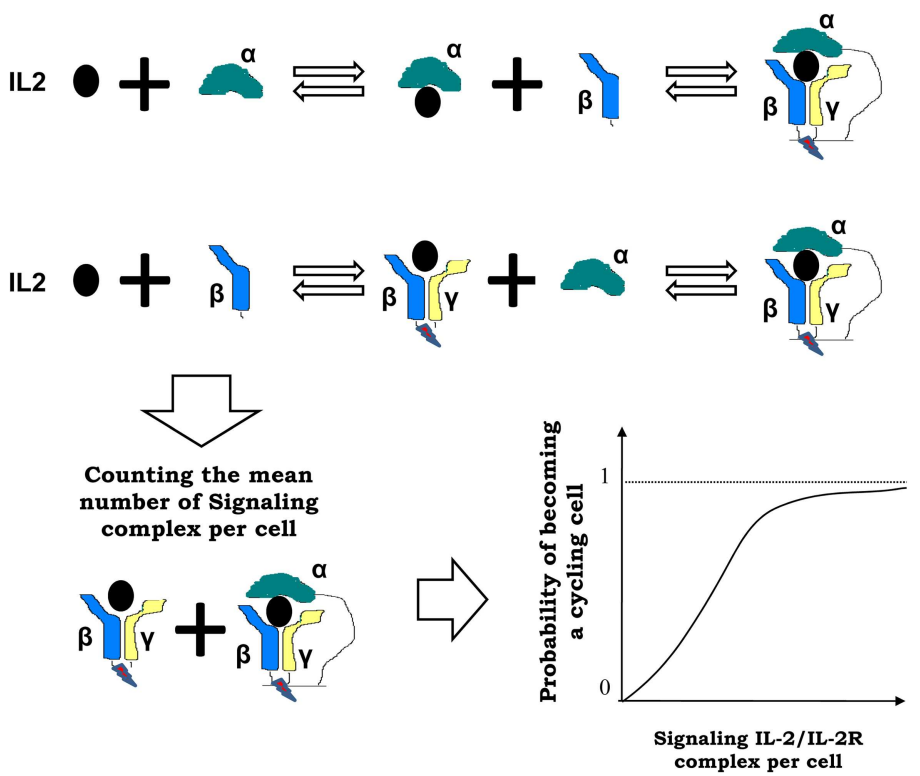


FIGURE 3 | Interactions between IL2 and T cells in the model are mediated by the IL2 receptor (IL2R), which is formed by three chains: alpha, beta, and gamma chain. These chains are combined dynamically in multi-step process at the cell surface, upon IL2 binding, to conform the two known signaling forms of the IL2 receptors: high affinity alpha-beta-gamma and intermediate affinity beta-gamma receptor. In the model, the mean number of such signaling IL2-IL2R complexes per activated T cell are counted, and the probability of becoming a cycling cell is computed with a sigmoid function of the mean number of bound cytokines signaling receptors per cell (as shown at the right side of the arrow).

antibodies, and immune-complexes) and the number of complexes IL2-IL2R and IL2-mAb-IL2R formed in a single cell membrane are described in "Dynamics of Molecules in the Lymph Node" in Appendix C.

SIMULATION OF DIFFERENT THERAPIES

Four types of treatments are simulated in the model: injections of IL2; injections of anti-IL2 monoclonal antibodies; injections of immune complex composed of a mixture of IL2 and anti-IL2 antibodies with a specified constant proportion of them; and injection of mutant variants of IL2.

Treatments are simulated to represent a continuous infusion of the involved molecules for a defined period of time. This is implemented by setting on, transiently, the external source term of the molecules involved in a specific treatment (i.e., IL2; IL2m; and/or anti-IL2 antibody). Two parameters always control treatment application: the "dose," which set up the total amount per day of IL2, IL2m, and/or anti-IL2 antibody infused; and the "treatment duration," which set the time period for which continuous infusion is maintained. In all cases, we explore how the dose and treatment duration determine the outcome of the system simulation. We study whether or not different treatments can condition a significant preferential expansion (dominance) of helper T cells or regulatory T cells or M cells in the system.

PARAMETER AND VARIABLE VALUES IN MODEL SIMULATIONS

Model parameters were previously calibrated in Ref. (27). The actual values of parameter used in our simulations are provided in **Tables 1–3**. The majority of the model parameters are fixed to values directly taken or derived from available independent experimental data; just a few parameters remain unknown, and their influence in result was explored inside a range of biologically reasonable values. Given the realistic values and units of the most model parameters used in the simulations, we report in this paper the values of treatments doses in milligrams and the values of treatment duration in weeks. However, the reader should note that our model is only roughly calibrated, thus one should believe on the order of magnitude and general qualitative trends of the predicted effects. But, the exact values of dose and treatment duration reported here to cause a given effect in the simulations should not be taken as a solid prediction.

The simulations of the model dynamics was implemented using the program Mathematica v.4.0.

RESULTS AND DISCUSSION

BASIC MODEL AND SIMULATIONS SETUP

The model is setup to study the basic homeostasis of the immune system of a mouse (27). Therefore the APCs in the model are interpreted as those APCs, which present self-antigens to T cells in the absence of infections. In consequence, the $CD4^+$ T cells in the model are taken to represent the populations of auto-reactive E and R cells, which significantly recognizes the existent self-antigens and thus interact with the available APCs.

Two main problems are then studied in the model simulations. (a) The basic dynamics states of the system in the absence of treatments; and (b) The effect of perturbations which represent specific IL2 modulation treatments on the stability of these dynamics states.

TOLERANCE AND IMMUNITY AS THE BASIC MODEL STEADY STATES (IN THE ABSENCE OF TREATMENT)

The model has two stable steady states which can be interpreted as natural tolerance and autoimmunity in the system. The steady state, which is interpreted as an autoimmune state (**Figure 4A**), is one where auto-reactive helper cells are significantly expanded while the auto-reactive Regulatory T cells are outcompeted from their cognate APCs. This steady state is also characterized by the existence of high levels of free IL2 and some subsequence expansion of the memory $CD8^+$ T cells population (M cells) in the lymph nodes. The steady state, which is interpreted as natural tolerance in the model, is one where the auto-reactive E and R cells co-exist in a dynamic equilibrium (**Figure 4B**). In this steady state the expansion of the auto-reactive helper cells is actively controlled by their interaction with the auto-reactive Regulatory T cells, the amount of free IL2 remains very low and the size of M cell population remains close to its basal homeostatic level.

A key dynamical property of the model is the existence of a parameter regime where these steady states of tolerance and autoimmunity can co-exist. This is a regime of bistable behavior (**Figure 4C**), where the model could evolve dynamically into either to the autoimmune or the tolerant steady state, but depending on the initial conditions used to seed the simulation without any change of parameter values (i.e., changing the initial proportion of auto-reactive E to R cells). The model is set to operate inside this bistable parameter regime. Thus in equilibrium, in the absence of treatments, the system will be either in the tolerant or the autoimmune steady state referred above. Such parameter choice is required to explain properly with the model the results of adoptive transfer experiments in mice, where transferring different proportions of $CD4^+CD25^-$ (helper) and $CD4^+CD25^+$ (regulatory) T cells into immune deficient mice (those lacking T cells, $Rag^{-/-}$ or $nu^{-/-}$), they either reconstitute a normal (tolerant to self-antigens) immune system or develop an autoimmune disease mediated by the uncontrolled expansion of the transferred auto-reactive $CD4^+$ T cells (28).

Moreover, it is important to note that the model reviewed here is an extension of the cross-regulation model of immunity, which studies the interaction of helper and regulatory $CD4^+$ T cells in the lymph node of the normal mice (33). Interestingly, despite of substantial increase on the number of variable and parameters, the new model conserves the three main dynamical properties of the original one reviewed in Ref. (34). In Ref. (28), three parameter conditions were presented as necessary in the extended model to behave as the original model and therefore to explain the same phenomenology. These conditions are:

(1) Regulatory T cells have to be more efficient using IL-2 at low concentrations than helper and memory T cells.
(2) The existence of a cytokine alternative to IL-2 that promote helper T cell proliferation and survival.
(3) The helper cells must become activated and proliferate more rapidly than Regulatory T cells in conditions of IL-2 excess.

A detailed discussion of the validity of these constrains, from an experimental point of view, is provided in Ref. (28).

Table 1 | Variables and parameters appearing in the equations that model the dynamics in the blood compartment.

Variables	Definitions	
$IL2_S$	Total number of IL2 molecules in the blood	
$IL2m_S$	Total number of IL2m molecules in the blood	
Ab_S	Total number of anti-IL2 mAb in the blood	
$IL2_S^{Ab}$	Total number of IL2-mAb complexes in the blood	
IL2	Total number of free IL2 molecules (non-conjugated to IL2R at the cell membrane) in the lymph node	
IL2m	Total number of free IL2m molecules (non-conjugated to IL2R at the cell membrane) in the lymph node	
Ab	Total number of free anti-IL2 mAb (non-conjugated to IL2-IL2R complex at the cell membrane) in the lymph node	
$IL2^{Ab}$	Total number of free IL2-mAb complexes (non-conjugated to IL2R at the cell membrane) in the lymph node	
Symbolic labels	**Definitions**	
j	Symbolic label that denotes the different IL2R chains: $j = \alpha$ (alpha chain) and $j = \beta$ (beta chain)	
l	Symbolic label that denotes the possible functional states of the T cells: $l = N$ resting state, $l = A$ activated state and $l = C$ cycling state	
Parameters	**Definitions**	**Values used in simulations**
Γ_i	External influx of IL2, typically used to simulate IL2 addition treatment	–
K_{pi}	Rate of IL2 production by helper $CD4^+$ T cells upon activation	10^3 M/h
K_{di}	Elimination rate of IL2 in the blood	Ln(2)/10 min
N_{LN}	Total number of equivalent lymph nodes considered in the system	10
D_{il}, D_{ab}	Diffusion rate for the exchange of IL2 and mAbs, between the blood and peripheral lymph nodes	10^{-7} L × Ln(2)/10 min
V_S, V_{LN}	Volume of the blood and lymph node compartments, respectively	2.5×10^{-3} L, 10^{-6} L
fve	Fraction of the lymph node volume, in which molecules and mAbs can diffuse	0.1
K_{on}^{Ab}, K_{off}^{Ab}	Association and dissociation constants of IL2-mAb complexes	Face alpha mAb: 1.5×10^5 $M^{-1}s^{-1}$, 1.4×10^{-4} s^{-1}; face beta mAb: 2.3×10^4 $M^{-1}s^{-1}$, $6.6 \times 10^{-5}s^{-1}$
Γ_{mi}	External influx of IL2m, typically used to simulate IL2 addition treatment	–
Γ_{ab}	External influx of mAb, typically used to simulate anti-IL2 mAbs addition treatment	–
K_{da}	Elimination rate of mAbs and IL2-mAbs complexes in the blood	Ln(2)/3 days
N_A	Avogadro's number	$6{,}02 \times 10^{23}$ mol^{-1}

RESPONSE TO TREATMENTS THAT MODULATE IL2 CONCENTRATION

In following sections, the effects of different treatments, which aim to modulate IL2 activity, are studied. Treatments simulate a continuous infusion for a defined period of time of the involved molecules (IL2, IL2m, and/or anti-IL2 antibody). Two parameters control their application: the "dose," which set up the total amount per day of IL2, IL2m, and/or anti-IL2 antibody infused; and the "treatment duration," which set the time period of sustained infusion. Treatments are always applied in a system which is initially set to a dynamic equilibrium (i.e., either into the tolerant or the autoimmune steady state). We systematically study, whether a given treatment induces a significant change in the initial proportion of Regulatory (R) versus Helper (E + M) T cells, both transiently or permanently. We interpret that a treatment promotes immunity when it induces a transition from the tolerant steady state (dominated by R cells) to the autoimmune steady state (dominated by E cells). We interpret that a treatment promotes tolerance when it induces a transition from the autoimmune state to the tolerant steady state.

Simulating the injection of IL2

Simulations of IL2 injections show that, when this treatment is applied to a system initialized into the autoimmune steady state, it is unable to take the system into the tolerant steady state, irrespectively of the dose and treatment duration chosen. Moreover, it further promotes the expansion of the auto-reactive E cells and the M cells (**Figure 5**) reinforcing transiently the ongoing autoimmune response. However, when this treatment is applied

Table 2 | Variables and parameters appearing in the equations that model the T cells dynamics.

Variables	**Definitions**	
E_N, E_A, E_C	Total number (conjugated plus non-conjugated) of resting, activated, and cycling E cells	
R_N, R_A, R_C	Total number (conjugated plus non-conjugated) of resting, activated, and cycling R cells	
M_A, M_C	Total number (conjugated plus non-conjugated) of activated and cycling M cells	
Intermediate variables	**Definitions**	
E_N^B, E_A^B, E_C^B	Number of resting, activated, and cycling E cells conjugated to APCs	
E_T^B	Total number of conjugated E cells: $E_T^B = E_N^B + E_A^B + E_C^B$	
E_N^F, E_A^F, E_C^F	Number of resting, activated, and cycling E cells non-conjugated to APCs: $E_I^F = E_I - E_I^B,\ \forall I \in \{N,\ A,\ C\}$	
R_N^B, R_A^B, R_C^B	Number of resting, activated, and cycling R cells conjugated to APCs	
R_T^B	Total number of conjugated R cells: $R_T^B = R_N^B + R_A^B + R_C^B$	
R_N^F, R_A^F, R_C^F	Number of resting, activated and cycling R cells non-conjugated to APCs: $R_I^F = R_I - R_I^B,\ \forall I \in \{N,\ A,\ C\}$	
M_A^B, M_C^B	Number of activated and cycling M cells conjugated to APCs	
M_T^B	Total number of conjugated M cells: $M_T^B = M_A^B + M_C^B$	
M_A^F, M_C^F	Number of activated and cycling M cells non-conjugated to APCs: $M_I^F = M_I = M_I^B,\ \forall I \in \{A,\ C\}$	
F	Total number of APC conjugation sites that remain free in the system	
SigE, SigR, SigM	Number of bound cytokines signaling receptors at the surface of an activated E, R, and M cells	
Symbolic labels	**Definitions**	
I	Symbolic label that denotes the possible functional states of the T cells: $I = N$ resting state, $I = A$ activated state, and $I = C$ cycling state	
Parameters	**Definitions**	**Values used in simulations**
Γ_e, Γ_r	Input rate of new resting self-reactive E and R cells from the thymus	2.5×10^4 cells/day
K_A^E, K_A^R	Activation rate for resting E and R cells conjugated to APCs	Ln(2)/2 h, Ln(2)/6 h
K_P^E, K_P^R, K_P^M	Division rate for cycling E, R, and M cells	Ln(2)/4 h
K_S^E, K_S^R	IL2 signaling-waiting rate for activated E and R cells	Ln(2)/2 h
K_S^M	IL2 signaling-waiting rate for activated M cells	Ln(2)/4 h
K_d^E, K_d^R, K_d^M	Death rate for free resting E and R cells, and free activated M cells	Ln(2)/1 week
A	Number of total APCs	2×10^5
s	Total number of conjugations site per APC	5
K^E, K^R	Equilibrium conjugation constants (K_{on}/K_{off}) for E and R cells to the APC conjugation sites	$K_{on} = 10^{-13}\,\mathrm{L\,s^{-1}\,cell^{-1}}$, $K_{off} = 6 \times 10^{-4}\,\mathrm{s^{-1}}$
K^M	Equilibrium conjugation constants (K_{on}/K_{off}) for M cells to the APC conjugation sites	$K_{on} = 10^{-13}\,\mathrm{L\,s^{-1}\,cell^{-1}}$, $K_{off} = 6 \times 10^{-3}\,\mathrm{s^{-1}}$
α_E, α_R	Fraction of activated E and R cells reverting to the resting state in the absence of cytokine related signal	0.95
h	Hill coefficient at the sigmoid response curve	4
S_E, S_R, S_M	Sensitivities thresholds for E, R, and M cells to cytokines signal	500

Table 3 | Variables and parameters related to dynamics of IL2, IL2R, and mAb complexes formation.

Variables	**Definitions**	
$C_j^{E_I}$, $C_j^{R_I}$, $C_j^{M_I}$	Number of IL2 molecules bound to *j* chain of IL2R, at the surface of the indicated T cell type	
$Cm_j^{E_I}$, $Cm_j^{R_I}$, $Cm_j^{M_I}$	Number of IL2m molecules bound to *j* chain of IL2R, at the surface of the indicated T cell type	
$CAb_j^{E_I}$, $CAb_j^{R_I}$, $CAb_j^{M_I}$	Number of IL2/mAb complexes bound to the *j* chain of IL2R, at the surface of the indicated T cell type	
T^{E_I}, T^{R_I}, T^{M_I}	Number of IL2 molecules bound to high affinity IL2R (alpha + beta), at the surface of the indicated T cell type	
Tm^{E_I}, Tm^{R_I}, Tm^{M_I}	Number of IL2m molecules bound to high affinity IL2R (alpha + beta), at the surface of the indicated T cell type	
Intermediate variables	**Definitions**	
$P_j^{E_I}$, $P_j^{R_I}$, $P_j^{M_I}$	Number of IL2R of *j* chain free (not bound to IL2), at the surface of the indicated T cell type	
SigE, SigR, SigM	Number of cytokines signaling receptors at the surface of an activated *E*, *R*, and *M* cells	
Parameters	**Definitions**	**Values used in simulations**
K_{off}^j, K_{on}^j	Dissociation and association constant of IL2 to the *j* chain of the IL2R	$K_{off}^{\alpha} = 0.6\ s^{-1}$, $K_{on}^{\alpha} = 10^7 M^{-1}s^{-1}$, $K_{off}^{\beta} = 3 \times 10^{-3}s^{-1}$, $K_{on}^{\beta} = 3.4 \times 10^6 M^{-1}s^{-1}$
f_j	Parameter that control the properties of different IL2m	10^{-3}, 10^3
N_j	Switch parameter setting if the mAb blocks (=1) or not (=0) the interaction of IL2 with the *j* chain of the IL2R	0, 1
ila_{E_A}, ila_{M_A}	Number of cytokine signaling receptors, at the surface of an activated *E* and *M* cells, bounds to an alternative cytokine (not IL2)	10^8, 10^7
Ra^{E_I}, Rb^{E_I}	Total number of alpha and beta chains of IL2R per E cells in the state *I*	$Ra^{E_N} = 10$, $Ra^{E_A} = 10^4$, $Ra^{E_C} = 10^3$, $Rb^{E_I} = 10^3$
Ra^{R_I}, Rb^{R_I}	Total number of alpha and beta chains of IL2R per R cells in the state *I*	$Ra^{R_N} = 10^4$, Ra^{R_A}=10^5, Ra^{R_C}=10^4, Rb^{R_I}=10^3
Ra^{M_I}, Rb^{M_I}	Total number of alpha and beta chains of IL2R per M cells in the state *I*	Ra^{M_I}=10, Rb^{M_I}=10^4
$K_{on}^{\alpha\beta}$, $K_{off}^{\alpha\beta}$	Association and dissociation rates for the interaction of free beta chain to preformed IL2/alpha chain complexes, at the T cell membrane	$K_{on}^{\alpha\beta}$=2.2 × $10^{-3}s^{-1}$, $K_{off}^{\alpha\beta}$= $3 \times 10^{-3}s^{-1}$
$K_{on}^{\beta\alpha}$, $K_{off}^{\beta\alpha}$	Association and dissociation rates for the interaction of free alpha chain to preformed IL2/beta chain complexes, at the T cell membrane	$K_{on}^{\beta\alpha} = 0.6 \times 10^{-2}s^{-1}$, $K_{off}^{\beta\alpha} = 0.6\ s^{-1}$
K_{in}	Internalization (degradation) rate of signaling IL2/IL2R complex by T cells	$K_{in} = 0.04\ min^{-1}$

to a system initialized in the tolerant steady state it reinforces the preexistent tolerance, by further expanding the regulatory populations (**Figure 5**). Interestingly, increasing the IL2 dose applied to a preexistent tolerant state could induce immunity by expanding the M cells; although, this effect is obtained for significantly high (unrealistic) values of the IL2 dose.

Thus overall in the model, IL2 injections appear to reinforce the preexistent steady state, this is expanding transiently either the R or the E cells respectively for a preexistent tolerant or autoimmune situation. A closer look to the model behavior qualitatively explains these results. Briefly: in a preexistent autoimmune steady state there is an excess of IL2 in the lymph node, thus is not lack of IL2 what limits regulatory T cell expansion, is their competition with auto-reactive E cells for the cognate APCs. In consequence injecting IL2 would never reestablish tolerance. In a preexistent tolerant steady state, there is a small amount of IL2 in the lymph node, which is almost exclusively used by the regulatory T cells, limiting their expansion. The helper T cells do not expand as result of the direct suppression of their activation exerted by the R cells. In this situation the injection of IL2, naturally leads to the enhanced expansion of R cells reinforcing the suppression over the E cells. Only when the IL2 concentration is extremely high at the lymph nodes it triggers a significant expansion of the Memory T cells, signaling through the intermediate affinity IL2 receptor beta-gamma. The excessive expansion of the M cells in the system affects the suppressive interaction between E and R cells at the APCs, since these cells, although much weakly, also interact with and compete for the available APCs.

Interestingly the latter model predictions are indeed compatible with existent experimental observations and further provide a guideline for its future practical application. On the one hand, the reinforcement of ongoing immune reactions by IL2 injections, predicted by the model, explains classical observations on *in vivo* animal models, where IL2 have been shown to potentiate immune reactions to viral infection (35) and to well-adjuvated vaccines (1–3). In these systems the immune response induced to

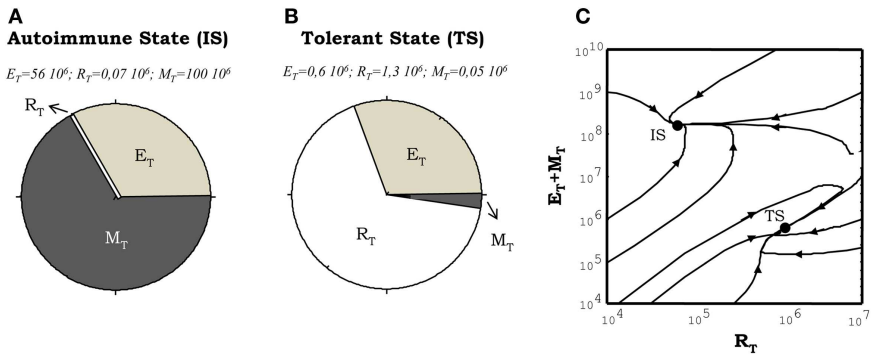


FIGURE 4 | Illustration of the steady states obtained from numerical simulations of the model. **(A,B)** shows the proportion of the total T cell number corresponding to helper (E), regulatory (R), and memory (M) T cells. The situation showed in **(A)**, corresponds to the autoimmune steady state (IS) where the memory and helper T cells dominate the system. The situation depicted in **(B)**, correspond to the tolerant steady state (TS) where the regulatory T cells dominate the dynamics. The graph in **(C)** illustrates how these two types of steady state of the system co-exist in the same region of parameter values (the region of bistability). It is shown how different initial conditions, changing just the proportion of E, R, and M cells at time $t = 0$, leads to trajectories taking the system either in the tolerant (TS) or the autoimmune (IS) steady state.

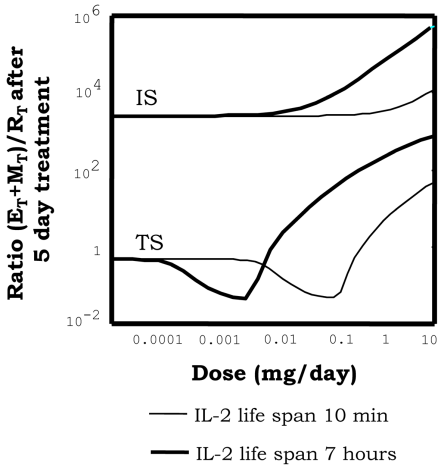


FIGURE 5 | Effect of injections of IL2, on the proportion of helper + memory T cells versus regulatory T cells [ratio (E + M)/R], in a system initialized either in tolerant (TS) or the autoimmune (IS) steady state. The graph shows the ratio (E + M)/R attained in the system right after 5 days of continuous injections of the indicated dose (*x* axis of the graph) of an IL2 with either 10 min (thin curves) or 7 h (thick curves) life span in solution. It can be seen how when the simulations start with a system at the TS, the ratio (E + M)/R reduce its values for intermediate dose of the treatment. This is a direct consequence of a preferential expansion of the R cells in the system. However, if the dose is further increased then the ratio (E + M)/R is significantly increased. This is a direct consequence of the expansion of helper and memory T cells, by the treatment application. When the treatment start on a system at the IS, then increasing the dose always leads to an increase of the ratio (E + M)/R. This is, it further increments the number of E and R cells in the system. Interestingly increasing the life span of the injected IL2 moves to lower values the dose ranges where treatment is effective, but does not change the qualitative pattern of response observed.

the involved foreign antigens, which are most probably loosely or just not controlled by regulatory T cell activity, is further promoted by the injected IL2. Furthermore, the observed enhancement of immunity, in these experimental systems, might not relay just on the model predicted expansion of helper $CD4^+$ T cells. It might also involve important direct effects of IL2 on memory $CD8^+$ T cell and/or NK cells, which are known to be relevant in many of these particular systems. In any case the model here, will further predict that optimal application of IL2 for the purpose of enhancing immunity, will be obtained when providing IL2 after the immune reaction have already started and never before, because some reminiscent of immune regulation might still exist and could be potentiated by the added IL2.

On the other hand, the capacity of IL2 addition to reinforce natural tolerance mediated by regulatory T cells, predicted by the model, explains as well several experimental observations. Particularly, it explains clinical data stating that regulatory T cells populations are significantly expanded, both in cancer (9, 36) and HIV (37) patients, treated with IL2. Such effect might be related to the poor efficacy observed in these clinical applications of IL2. Particularly, in the case of cancer, less than 20% of the treated patients show some anti-tumor effect, perhaps, according to the model here, because just an small fraction of the patients, happen to have a naturally preexistent immune response against tumor antigens, which could be further enhanced by the injected IL2. In the case HIV patients, IL2 based therapy have led to the recovery of $CD4^+$ T cells counts, but the patients do not seem to recover their capacity to fight general infections, perhaps, according to the model here, because this treatment is just reinforcing tolerance mediated by regulatory T cell activity.

Furthermore, this second model prediction also explains many results in pre-clinical animal models. It explains, for instance, that IL2 injections can prevents allograft rejection (10); or attenuate the induction of Experimental Autoimmune Encephalomyelitis (EAE) (10); or fully prevent the development of diabetes in the NOD mice (11). Interestingly, in the EAE and allograft reaction models the latter effects are observed for scheme of IL2 applications where this cytokine is injected in the system before implanting the allogeneic tissue or before inducing the EAE. This is before the immune/autoimmune reaction has been expanded; i.e., when

there is a preexistent natural tolerance, mediated by regulatory T cells, which could be reinforced by the applied treatment. However, in the NOD mice model, recent data (12) have shown that IL2 treatment at the onset of diabetes could revert disease development. Interestingly, in this "therapeutically relevant scenario" treatment efficacy is much lower than in the preventive settings. Only 40–60% of the NOD mice appear to be cured, while 100% of the NOD mice are diabetes free when treating in the preventive settings. Whether or not at the onset of NOD mice diabetes the balance between regulatory T cells and helpers T cells have been fully disrupted in favor of immunity, just as considered in our model simulations of an autoimmune disease therapeutic scenario, is a matter of discussion. Actually Grinberg-Bleyer et al. have shown that at the onset of NOD diabetes a significant amount of regulatory T cells can still be found in the pancreas and its draining lymph node. Unfortunately in the NOD mice, the acute nature of diabetes development (with a full irreversible destruction beta islet) invalidates any displacement of the treatment application toward a more advanced stage of the disease, to better compare with our model predictions.

Simulating the injections of different anti-IL2 mAbs

Anti-IL2 antibodies are molecules that form complexes with the IL2, blocking or not its binding to the different chains of the IL2 receptor at the T cell surface and therefore interfering with the associated signaling process. Three classes of antibodies are systematically explored in our simulations following its documented existence in the literature (20, 38): (1) The anti-IL2 mAbs, which bind and thus block the site in the IL2 surface implicated on the interaction with the alpha chain of the IL2 receptor (referred here as the face alpha mAbs); (2) The anti-IL2 mAbs, which bind and thus block the site in the IL2 surface implicated on the interaction with the beta chain of the IL2 receptor (referred here as face beta mAbs); and (3) the anti-IL2, which block the binding of IL2 to all chains of the IL2 receptor (referred here as a fully blocking mAbs).

The injection of monoclonal antibodies anti-IL2, in the model simulations, when applied to a previously tolerant system could induce a breakdown of tolerance (**Figure 6A**), with the consequent transition of the system to the autoimmune steady state. Such effect can be obtained with the three classes of anti-IL2 mAbs studied, but it requires a minimal effective dose of the anti-IL2 mAb and treatment duration (**Figure 6A**) which varies significantly with the type of mAbs used. Face alpha mAbs are significantly more efficient than fully blocking or face beta mAbs (**Figure 6A**) in this simulation. Moreover, for the three classes of mAbs studied the higher the affinity for the IL2 the higher their efficacy in these simulation (27).

The effect of treatment with anti-IL2 mAbs in a system with a preexistent autoimmune reaction is also quite significant. In this case, the treatment is capable of resetting the system into the

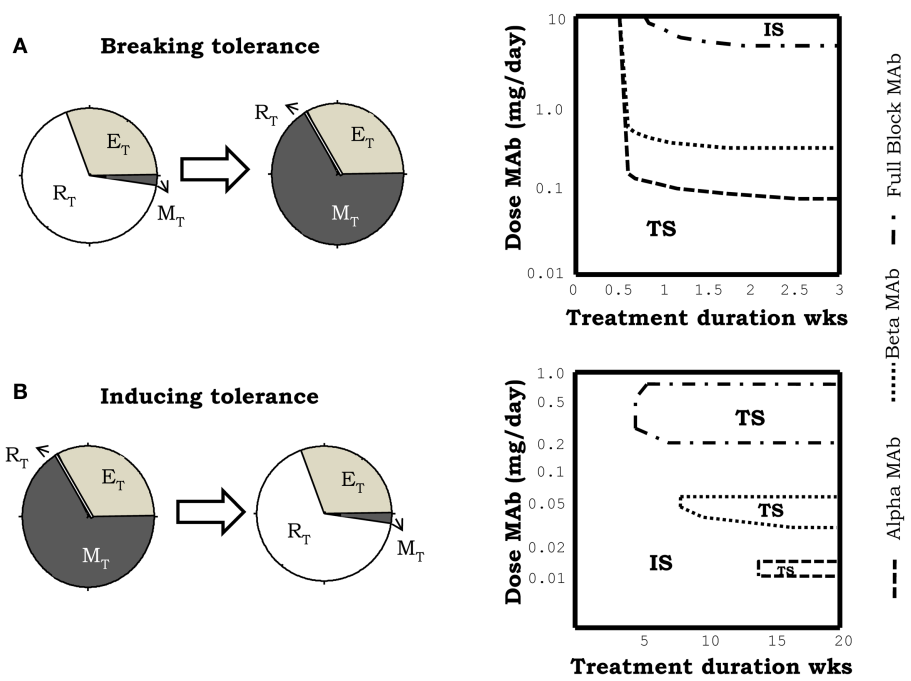


FIGURE 6 | Effect of the simulation of treatments of IL2 depletion, using different anti-IL2 antibodies. mAbs in the simulation, can block the interaction of IL2 with the alpha (face alpha mAb), or with the beta (face beta mAb) or with both (fully blocking mAb) chains of the IL2R. The graphs in **(A)** corresponds to the case in which the treatment induces a breakdown of the preexistent tolerant steady state, i.e., a transition to the autoimmune steady state. The graphs in **(B)** corresponds to the case in which treatment induces tolerance, taking into the tolerant steady state, a system initially set in the autoimmune steady state. Breakdown of a preexistent tolerant state requires a minimal effective dose of mAb and treatment duration [graph in the right side of **(A)**]. In this scenario, face alpha mAbs are more efficient than face beta or fully blocking mAb. This means that the latter need higher doses of mAb to achieve a similar effect. Induction of tolerance requires minimum treatment duration with a mAb dose inside an intermediate window of values [graph in the right side of **(B)**]. This effect is obtained when face alpha, face beta, or fully blocking mAbs are used.

tolerant steady state (i.e., inducing tolerance) (**Figure 6B**). This effect occur under quite restrictive treatment conditions: there is a minimal treatment duration required and the dose of the anti-IL2 mAbs used has to be set inside some particular intermediate range of values (**Figure 6B**). The tolerogenic effect of the anti-IL2 mAbs is obtained with all type of mAbs (**Figure 6B**). Differences in the mAbs affinity and mAbs class strongly change the dose range where this effect is observed[2].

Overall the simulations of IL2 depletion treatments using anti-IL2 antibodies predict that this type of therapy is able to break a preexistent tolerant state, inducing an autoimmune response, or to render tolerant a preexistent autoimmune system. A closer look to the model behavior qualitatively explains these results as follows. The injected mAbs appear to sequester the IL2, limiting its availability to provide signal to the T cells. When the treatment is applied into an initially tolerant steady state, the initial effective concentration of IL2 is low and it is further reduced to insignificant levels, where this cytokine is incapable to signal neither to E, R, or M cells. Therefore, if the treatment is sustained long enough, the number of R cells fall down to a minimum determined by the size of the thymic output, because the proliferation and survival of R cells is strictly dependent on IL2. But the number of E cells, on the other hand, set back to a value determined by the availability of the homeostatic cytokine of ILα, which they could use as alternative to IL2 signal. Therefore once the injected mAbs are cleared, the auto-reactive E cells could have some initial advantage in respect to the R cells, leading the T cell expansion, which drive the system into the autoimmune steady state. However, when the treatment is applied to an initially autoimmune system, the effective concentration of IL2 is quite large and it is reduced by the presence of the antibody. The efficacy of the mAbs to affect IL2 signaling in the different T cell population is strongly dependent on its affinity for the IL2 and the side of the IL2 recognized. For a very high antibody dose, the effective IL2 concentration falls to negligible values, which as before are unable to signal neither to E, R, or M cells. Thus the size of the auto-reactive E cell population is reduced to the value set by the availability of ILα and the number of R cell remains low in a value determined by the size of thymic output. When the injected antibody is cleared the system could return back to the autoimmune equilibrium. However, for some intermediate doses of the antibody, the effective IL2 concentration is reduced to values where it is unable to signal on the E and M cells, but it is still significant for the R cells, which are more sensitive due to their higher expression of the alpha chain of the IL2 receptor. Therefore, for these mAbs doses the E cell population is reduced to the minimal size, which can be sustained by the available ILα. But the R cells are stimulated to grow forcing the system to switch into the tolerant steady state.

[2]Treatment using the face alpha mAbs is the best option (it work for lower MAb dose windows, example shown in **Figure 6B**) when the mAb used has an affinity for IL2 lower than 10^{10} M^{-1}. But the capacity of this mAb to revert an autoimmune state is completely lost if the mAb affinity for IL2 is assumed to be higher. The face beta mAbs seems to work well for a larger range of mAbs affinities. Its effect is lost only for unrealistically high affinities (larger than 10^{11} M^{-1}) and it is always better than the one obtained with a fully blocking mAb. In Ref. (27), **Figure 7**, we presented results considering a mAb affinity higher than 10^{10} M^{-1}, where the face alpha mAb is not effective in reverting an autoimmune state.

The model prediction of a higher efficacy of treatments with face alpha mAbs, to break a preexistent tolerant steady state, relates to the impact of this type of mAbs on the dynamics of the M cells. Face alpha mAbs bind the available IL2 forming immune-complexes that can still signal through the intermediate affinity IL2 receptor (beta + gamma chain). This form of the receptor is prevalent in the M cells, thus face alpha mAbs partially redirect IL2 signaling into the M cells expanding this population. The growth of the M cells interferes with the dynamics of $CD4^+$ T cells, i.e., M cells consume the available IL2 and reduce the capacity of $CD4^+$ T cells to interact with the APCs. The combination of the latter effects explains the advantage of the face alpha mAb to break a pre-existent tolerant steady state. On the other hand, the differences observed between fully blocking and face beta mAbs in the model simulations (compare dose dependencies in **Figure 6**), must rely on the fact that face beta mAbs do not block the interaction with the alpha chain of the IL2 receptor, conditioning the attachment of the immune-complexes formed to cells that express this molecule at the cells surface. These interactions significantly alter the bio-distribution of both the IL2 and the antibody.

Interestingly, the latter model predictions are indeed compatible with existent experimental observations. On the one hand, the predicted capacity of treatments blocking IL2 activity to promote autoimmunity/immunity, explains observations where monoclonal antibodies against IL2 have been shown to promote effective immune responses to tumors (16) and to induce autoimmune disease in naïve mice (13). In both cases, the model explains the observed effects as being associated to the treatment capacity to weaken regulatory cell activity, just as argued by their original authors. It must be also noted that in these reports the S4B6 anti-IL2 mAbs was used, a mAb which has been recently proven to block only the interaction of IL2 with the alpha chain of the IL2 receptor (38).

On the other hand, the model predicted capacity of IL2 blocking therapies to reestablish tolerance in the context of ongoing immune/autoimmune reactions, is not documented in the literature. This model prediction is very interesting from the practical perspectives for the treatment of autoimmune diseases. However, the fact that the predicted treatment effect just occurs for a particular intermediate range of antibody doses, applied during a relatively long period of time, makes difficult the practical implementation of the treatment. To overcome the latter problem we suggested, based on model simulations, an alternative/simpler strategy to capitalize this therapeutic effect. A large initial dose of the mAb could be used, reducing it periodically with a fixed rate. With this alternative strategy the model predict a much simpler dose dependency (see **Figure 7**) of treatment efficacy, i.e., the applied initial dose used must be large enough (as to induce significant initial immunosuppression), and the reduction rate used should be sufficiently slow.

Simulating the injection of IL2/mAb immune-complexes

Immune-complexes of IL2 plus anti-IL2 mAbs (in a 1:2 mAb:IL2 molar proportion), has been recently highlighted as a novel therapeutic strategy (18, 20, 39) which could significantly potentiate the activity of the IL2 *in vivo*. Intuitively it should be expected that the therapy with immune-complexes share properties with

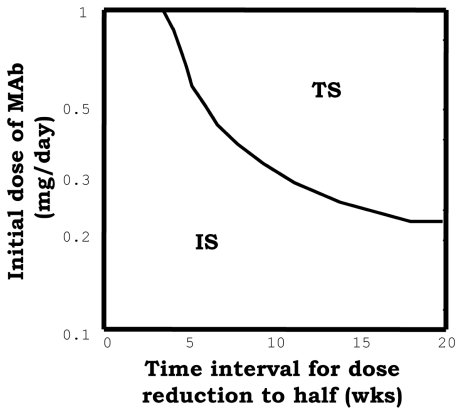


FIGURE 7 | The graph summarizes the results of simulations of a model system set initially to the IS and then perturbed with, an initial dose of the face beta mAb, which is periodically reduced to the half at the indicated time (*x* axis). The curve indicates the minimal value of the initial mAb dose required to induce tolerance (taking the system into the tolerant steady state) with the applied treatment.

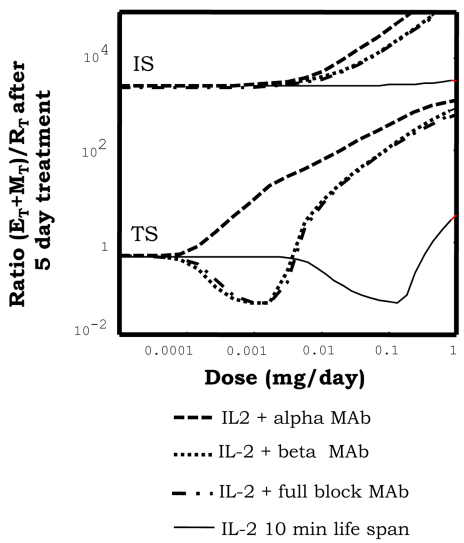


FIGURE 8 | Effect of injections for five days of the indicated doses of immune-complexes of IL2 plus antibodies anti-IL2, on the ratio of helper + memory T cells versus regulatory T cells [ratio (E + M)/R], in a system initialized either in tolerant (TS) or the autoimmune (IS) steady state. Different immune-complexes differ on the class of mAb used to form it (face alpha, face beta or fully blocking mAbs). immune-complexes are always formed with a 1:2 molar ratio of mAbs:IL2 and the dose applied is reported in terms of the mass of IL2 injected. If the simulations start with a system at the TS, immune-complexes formed with face beta or fully blocking mAbs reduce the ratio (E + M)/R for some intermediate dose values and then increases it for higher dose values. This is a pattern of response, qualitatively similar to that obtained with IL2 injection, but significantly displaced to the range of lower doses of IL2. If face alpha mAbs are used to form the complex the pattern of response obtained is qualitatively different. The ratio (E + M)/R always increase (favoring the expansion of E and M cells) and the larger the dose applied the larger the increment. If the simulations start with a system at the IS, all the possible immune-complexes behave qualitatively like the IL2 alone, they promote in dose-dependent way a further expansion of E and R cells, increasing the ratio (E + M)/R.

the therapies based on its basic components, but their comparative efficacy shall depend on the class and the affinity of the mAbs used.

In our simulations, immune-complexes can either reinforce or weaken a preexistent tolerant steady state depending on the class of mAb used on its formulation. **Figure 5**, shows how the injection of immune-complexes significantly changes the number of E, R, and M cells in the system initially set to the tolerant steady state. Immune-complexes formed with beta face or fully blocking mAbs induce a quite significant transient expansion of Regulatory T cells (reinforcing the tolerant state). This transient expansion of the Regulatory T cells is significantly larger than the one induced by an equivalent treatment with IL2 alone and it is maximal for mAbs with some intermediate affinity for the IL2 (27). However, immune-complexes formed with face alpha mAbs have a quite different effect in these simulations (**Figure 8**). They could also expand the R cells, but they expand much more in comparison the M cells in the lymph node. The capacity of this immune complex to expand M cells became larger as their affinity for the IL2 is increased (27). Interestingly for the three classes of immune-complexes a sufficiently high dose of the latter treatment could induce a breakdown of tolerance. But only immune-complexes based on face alpha mAbs perform better in this task than the therapy based on the anti-IL2 mAb or the IL2 alone (**Figure 9**).

When applied to initially autoimmune steady states, all immune-complexes fail to reestablish tolerance steady state. As the injection of IL2 the immune-complexes further reinforce a preexistent autoimmune steady state, expanding the helper and memory T cells (**Figure 8**).

Summarizing the results above shows that immune complex can sometimes synergistically potentiate the effects of IL2 and mAbs. Complexes based on face alpha mAbs do promote immunity primarily by expanding the M cells, and leading ultimately to a quite efficient breakdown of a preexistent tolerant steady state. Complexes based on face beta mAbs, can efficiently reinforce tolerance expanding significantly the R cells preexistent in the tolerant steady state. Face alpha mAbs for immune-complexes are better with the highest possible affinity, but face beta mAbs could be better with some intermediate affinity values.

Qualitatively the effects of immune-complexes can be explained based on two main dynamical properties in the model: (A) In the immune complex the IL2 is protected from degradation. While bind to the mAbs the IL2 has a life span of 3 days (like the mAbs), which is significantly larger than the life span of 10 min reported for free IL2. (B) Immune-complexes block different sites in the surface of IL2 conditioning its preferential interaction with different cell populations, accordingly to their differential expression of the IL2 receptor chains. Face alpha mAbs, form immune-complexes that bind and signal through the beta + gamma pair of IL2 receptors. Thus, since beta chain is over-expressed by the M cells, this complex preferentially redirect the IL2 signal to these cells. Following this analysis one could easily explain why this type of immune complex has a maximal efficiency when the affinity of the face alpha mAbs used is high. With high affinity mAbs, the IL2 is more protected from degradation, and the signaling is maximally

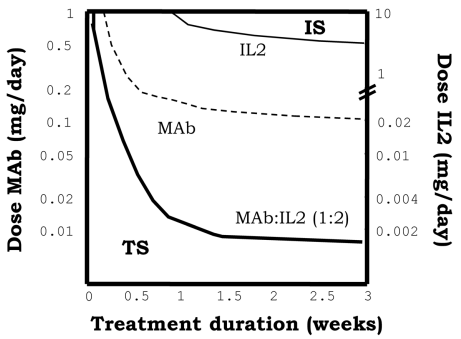


FIGURE 9 | The graph shows the minimal effective dose of mAb (left *y* axis) or IL2 (right *y* axis) versus treatment duration, required to induce the transition to the IS in a system initialized in the TS, for the treatment with immune-complexes formed with face alpha mAbs in the optimal molar proportion 1:2 (mAb:IL2). For direct comparison the equivalent curves obtained for treatments with the same mAbs alone or the IL2 alone are also depicted. It can be seen that the injection of this class of immune complex is more efficient than the injection of the mAb or the IL2 alone to breakdown tolerance in an initial tolerant system, i.e., it requires less dose of either the mAbs or IL2 as compared to the independent treatments.

redirected to the M cells. Face beta mAbs form immune-complexes unable to signal in any class of IL2 receptor. Thus to mediate any biological activity this type of complex has to partially dissociate, working as a controlled source of free IL2. If the affinity of the face beta mAbs in the complex is too high then the IL2 is never released and the immune-complexes have no effect at all. If the affinity of the face beta mAbs is too low, then injecting the complex is like injecting IL2 alone. However, if the affinity of the face beta mAbs in the complex is larger than the affinity of the dimeric IL2 receptor (beta + gamma chain), but lower than the affinity of the trimeric IL2 receptor (alpha + beta + gamma chain), the IL2 in the complex is easily release to provide signal through the high affinity trimeric IL2 receptor, but not through the intermediate affinity dimeric IL2 receptor. In this way the face beta based immune-complexes provided a preferential signaling to the regulatory cells, which overexpress the alpha chain of the IL2 receptor.

Interestingly the model results explain available pre-clinical data on the use of immune-complexes of IL2-anti-IL2 mAbs. Our observations that immune-complexes formed with face alpha or face beta mAbs expand different cell populations when injected *in vivo* into a normal (tolerant) mouse are fully compatible with the results reported in Ref. (18, 20, 39). In these experiments, the S4B6 mAb (a face alpha mAbs) is shown to form immune-complexes that strongly expands $CD8^+CD44^+$ T cells and to a lesser extent the R cells (20). This face alpha immune complex has been also used to increment the immune response induced with a vaccine (17), showing a significantly higher efficiency than IL2 alone. Moreover, the group of Jonathan Sprent have shown that JES6-1 (originally described as a face beta mAbs (20), although it has been recently observed that it also blocks the interaction with CD25, behaving more like a fully blocking mAbs) induce a larger expansion of Tregs ($CD4^+CD25^+Foxp3^+$ T cells) than the injection of IL2 alone in the same experimental setting (20). Remarkably, this latter type of immune complex has been shown to reinforce a preexistent tolerance state, preventing graft rejection or autoimmune disease induction in mice (18). But it showed no effect when applied in a therapeutic setting, this is just after the onset of the autoimmune disease or the initiation of skin graft rejection process (18).

The simulations, however, propose some interesting guidelines to improve the therapeutic effect of immune-complexes. They predict that in the case of complexes using face alpha mAbs, the best strategy is to use mAbs with the higher affinity available. But in the case of immune-complexes formed with face Beta or fully blocking mAbs, the use of intermediate affinity mAbs is recommended. Other important prediction of our model simulations is that treatment with immune-complexes based on face beta or fully blocking mAbs are useful to reinforce a preexistent tolerant state preventing the induction of autoimmunity, but it would be quite inefficient to therapeutically treat an already established autoimmune disorder. For the later task, the best strategy would be to use the anti-IL2 mAbs alone following the strategies described in Section "Simulating the Injections of Different Anti-IL2 mAbs."

Simulating the injection of IL2 mutants

Several mutant variants of IL2 have been designed aiming to improve the therapeutic efficacy of wild-type IL2 in cancer therapy. Most strategies, so far explored, involve the development of IL-2 variants with an either reduced or increased binding affinity for the alpha or the beta chain of the IL2R. In this section three particular classes of mutants are simulated: (a) IL2 Mutant with a reduced conjugation affinity for the alpha chain of the IL-2R as the one described in Ref. (40) (referred here as No-alpha mutants); (b) IL2 Mutant with an increased conjugation affinity for the alpha chain of IL-2R as the one described in Ref. (41) (referred here as Alpha-Plus mutants); (c) IL2 Mutant with an increased affinity for the beta chain of the IL2R as the one described in Ref. (42) (referred here as Beta-Plus mutant). These three classes of IL2 mutants provide a functional IL2 signal to T cells, since they keep binding beta and gamma unit of the IL2 receptor (i.e., they are IL2 agonists). But they might be expected to alter the natural balance in which wild-type IL2 is consumed by different T cell types, resulting on a significantly different overall dynamics.

Figure 10A show the effect of injecting different IL2 variants in a system initially set in the tolerant steady state. As described in Section "Simulating the Injection of IL2," injection of wild-type IL2 transiently reinforce this preexistent tolerant steady state, preferentially expanding the Regulatory T cells in the system. Alpha-Plus IL2 mutants, exhibit a similar response pattern than wtIL2, but with an even more marked preferential expansion of the regulatory T cells. In contrast, No-Alpha and Beta-Plus mutants show a completely different response pattern than wild-type IL2. This class of mutants expand preferentially the helper T cells (E + M), rather than the regulatory T cells at all injection doses. Moreover injections of the three classes of mutants, as for the wild-type IL2, could lead to a breakdown of tolerance in the system when the dose used is significantly increased. However, the minimal dose required for such effect is significantly lower for the No-Alpha mutants and Beta-Plus mutant **Figure 10B**, than for wild-type IL2 and Alpha-Plus mutants.

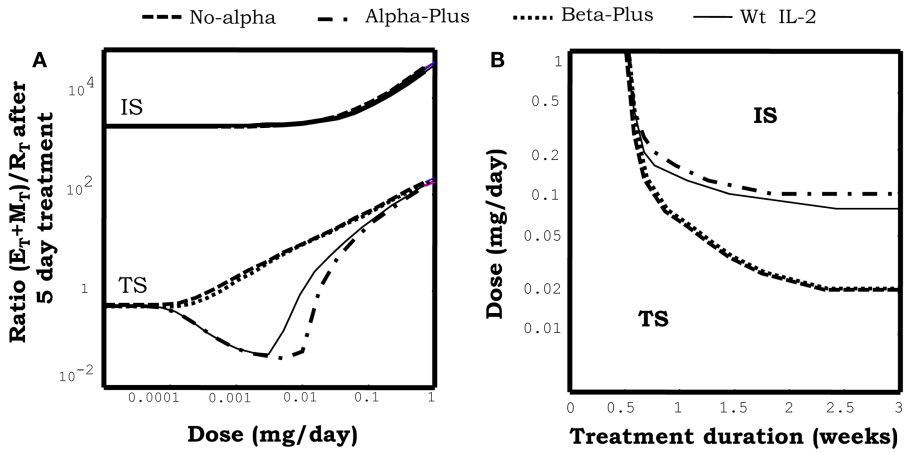


FIGURE 10 | The graph in (A) shows the effect of injections for 5 days at the indicated dose of different mutant variants of IL2 on the ratio of helper + memory T cells versus regulatory T cells [ratio (E + M)/R], in a system initialized either in the tolerant (TS) or the autoimmune (IS) steady state. Mutants differ on their capacity to bind to the different chains of the IL2R. Alpha-plus and Beta-plus mutants have higher binding affinity than wild-type IL2 respectively for the alpha chain ($f_\alpha = 1000$, $f_\beta = 1$) and the beta chain ($f_\alpha = 1$, $f_\beta = 1000$), while No-alpha mutant lack the binding to the alpha chain ($f_\alpha = 0.001$, $f_\beta = 1$). All mutant variants were simulated with a life span of 7 h. If the simulations start with a system at the TS, Alpha-plus mutant reduces the ratio (E + M)/R for some intermediate dose values and then increases it for higher dose values. This is a pattern of response, qualitatively similar to that obtained with IL2 injection, although with a slighter wider range of treatment dose with ratio (E + M)/R reduced from its starting value. If No-alpha or Beta-plus mutants are used the pattern of response obtained is qualitatively different. The ratio (E + M)/R always increase (favoring the expansion of E and M cells) and the larger the dose applied the larger the increment. If the simulations start with a system at the IS, all the mutants variants behave like the wild-type IL2, they promote in dose-dependent way a further expansion of E and R cells, increasing the ratio (E + M)/R. The graph in **(B)** shows the minimal effective dose versus the treatment duration, required to induce the transition to the IS in a system initialized in the TS, for the treatment with different variants of IL2 mutants. It can be seen that the injection of No-alpha and Beta-Plus IL2 mutants is more efficient than the injection of wild-type IL2 alone to breakdown tolerance in an initial tolerant system, i.e., it requires less dose to achieve a similar effect.

Figure 10A also shows the effect of injecting different classes of IL2 mutants in a model system initially set in the autoimmune steady state. In this case the three classes of mutants behave quite similarly to wild-type IL2, i.e., None of them is able to promote a transition to a tolerant steady states, at any dose and treatment duration. Moreover, they reinforce the preexistent autoimmune steady state, further expanding the Helper and Memory T cells.

Overall the result in this section show that No-Alpha and Beta-Plus IL2 mutants behave quite similarly, being significantly better than wild-type IL2 to promote immunity. While Alpha-Plus mutants could be slightly better that wild-type IL2 to reinforce a preexistent tolerant state, expanding more the regulatory T cells. Qualitatively, the latter results could be easily understood in the model, by taking into account the differential expression of the high affinity/trimeric form (alpha + beta + gamma) and intermediate affinities/dimeric form (beta + gamma) of the IL2 receptors on the different T cell populations. Regulatory T cells, relay on the overexpression of the alpha chain of the IL2R, to have the highest expression of the high affinity form of the IL2R. Memory T cells relay in the overexpression of the beta chain of the IL2 R to have the higest expression level of the intermediate affinity form of the IL2R. The No-Alpha and Beta-Plus mutants have a similar impact in the balance of use of IL2 related signal in the model. In both cases the resulting mutants lack the preferential capacity to signal over Regulatory T cells at low concentration, which is characteristic of the wild-type IL2. Furthermore they will preferentially redirect the signal toward the memory T cells, and strongly promote immunity. As the reverse case the Alpha-Plus mutant, reinforce the capacity of the wild-type IL2 to signal preferentially over the Regulatory T cells, resulting on a better tool to reinforce a preexistent tolerance state.

The results obtained above are compatible with existent experimental data. Both No-alpha (40) and Beta-Plus (42) mutants have been shown to induce a more potent anti-tumoral response than wild-type IL2 in several transplantable tumor models in mice. The dynamic effects predicted *in silico* for these types of IL-2 mutants is qualitatively similar to those described in Sections "Simulating the Injection of IL2/mAb Immune-Complexes" for treatments with immune-complexes of IL-2 and anti-IL-2 mAbs, when face alpha mAb are used. Indeed No-Alpha mutants can be easily conceptualized as an extreme case of such immune-complexes if the affinity of the mAbs for the IL2 tends to infinity. From a quantitative point of view, in the model, IL2 mutants can become as efficient as the immune-complexes, only when its life span is set to be greater than 24 h. If the life span of the mutant is taken to be of 7 h (the one used in **Figure 10**), which is the one reported for a wtIL2 fused to a constant region of IgG (43), then one might need around 5–10 time more mutant than wtIL2 in the immune-complex to obtain an equivalent effect. However, since immune-complexes work *in vivo* at very low concentrations of IL2, 1–2 micrograms in mice (20), a quite reasonably small amount of the IL2 mutants would be required to induce a similar effect. Thus, these IL2 mutants can be useful tools to promote immunity, for instance to treat tumors or

to enhance the response to cancer vaccines. It is important to note that mutants might have some regulatory/developmental advantages as potential drugs in comparison to the immune-complexes, given the fact that they are single molecules.

The predicted capacity of Alpha-Plus mutants to reinforce pre-existent tolerant steady state, expanding the regulatory T cells, has never been evaluated. A mutant variant of IL2 with 1000 times' higher affinity for the IL-2Ra was developed by Rao et al. (41). But only the *in vitro* effect of this mutant on different T cell lines was evaluated. Potentially the Alpha-Plus mutant could be used to treat patients that would receive an organ transplant to reduce the risk of graft rejection. However from a quantitative point of view the mutant efficacy expanding regulatory T cells is predicted only as slightly better than that of wild-type IL2. Moreover, it is quite similar to that obtained with immune-complexes of IL2-anti-IL2 mAbs formed with face beta or fully blocking mAbs (see Simulating the injection of IL2/mAb immune-complexes), when its life span is set to be of around 7 h (the value used in **Figure 10**). This is when its life span is similar to that reported for a wtIL2 fused to a constant region of IgG (43). Therefore this Alpha-plus mutant or just simply the wild-type IL2 fused to Fc of IgG, could be a reasonable drug to prevent allograft rejection. They might have a similar effect to that reported for immune-complexes in mice, but being much simpler drugs to develop. Indeed a version of IL2 fused to Fc portions of immunoglobulin is already available (43).

CONCLUDING REMARKS

Mathematical modeling of the IL2 and T-cell dynamics, considering the dual role of IL2 in its interaction with regulatory and helper $CD4^+$ T cells, is able to explain the complexity observed in the effects of IL2 modulating treatments. In this sense, we show that the model explains a large amount of available clinical and pre-clinical data. Moreover, it predicts optimal strategies for the future application of these treatments:

(A) Mutant variants of IL2, either with reduced affinity for CD25 (the alpha chain of IL2 receptor) or an increased affinity for CD122 (the beta chain of IL2 receptor), and with an increased life span in circulation (for instance fusing them to Fc portion of IgG), are the best strategy to potentiate immunity alone or in combination with vaccines.

(B) Increasing IL2 life span in circulation, either by fusing it with larger proteins or forming complexes with mAbs that block the interaction of IL2 and CD122 (the beta chain of the IL2 receptor), significantly potentiate its capacity to reinforce a preexistent natural tolerance, further expanding the regulatory T cells. This effect might be useful to treat patients that would receive an organ transplant, reducing the risk of graft rejection.

(C) Anti-IL2 antibodies which block the interaction of IL2 with CD122, CD25, or both can be used to treat an ongoing autoimmune disorder, promoting the induction of tolerance. The best schedule for this therapy is to start treatment with a high dose of the mAb (one capable to induce some immune suppression) and then scale the dose down slowly the dose in subsequent applications.

Last, but not least, it is important to highlight that our model has focused on the control that IL2 exerts on T cell cycle progression, impacting both in T cell proliferation and survival. We have neglected some other reported roles of IL2 in T cell differentiation. For instance, IL2 has been reported to increase the suppressive capacity of the Regulatory T cells (12); to condition the differentiation of CD8 T cells into a memory phenotype (44, 45); to induce together with TGFb, the generation of the so called induced Tregs from naïve $CD4^+$ T cells (46). We believe these phenomena, although important in some experimental contexts, are not essential to understands the main phenomenology studies in this paper. In future studies, the current model could be extended to include some or all of the above referred interactions of IL2.

Moreover, severe toxicity, i.e., the appearance of the cytokine storm and the vascular leak syndrome, is perhaps the major limitation known today of the practical application of IL2 modulation treatments in clinics. Our model cannot be used to simulate directly the toxic effects of the different IL2 modulation treatments studied. It could only be used to predict strategies that optimize the expected therapeutic efficacy related to the balance between regulatory and effector $CD4^+$ T cells. However, a recent report by the group of Boyman (47) has shown that vascular leak syndrome, which leads to severe pulmonary edema, is caused by the direct interaction of IL2 with its high affinity receptor expressed in lung epithelial cells. They demonstrated that treatment with immune-complexes of IL2 + S4B6 mAbs (anti-IL2 mAb which interferes the binding of IL2 to the alpha chain of IL2 receptor), prevents vascular leak syndrome while inducing a potent anti-tumor response. Furthermore, in Carmenate el al. (40), treatment with IL2 mutants with a reduced affinity for CD25 (no-alpha mutant) was shown to be less toxic than treatment with wild-type IL2. These experimental observations support the practical feasibility of some of our model predictions.

REFERENCES

1. Kudo-Saito C, Garnett CT, Wansley EK, Schlom J, Hodge JW. Intratumoral delivery of vector mediated IL-2 in combination with vaccine results in enhanced T cell avidity and anti-tumor activity. *Cancer Immunol Immunother* (2007) **56**(12):1897–910. doi:10.1007/s00262-007-0332-1
2. Fishman M, Hunter TB, Soliman H, Thompson P, Dunn M, Smilee R, et al. Phase II trial of B7-1 (CD-86) transduced, cultured autologous tumor cell vaccine plus subcutaneous interleukin-2 for treatment of stage IV renal cell carcinoma. *J Immunother* (2008) **31**(1):72–80. doi:10.1097/CJI.0b013e31815ba792
3. Lin CT, Tsai YC, He L, Yeh CN, Chang TC, Soong YK, et al. DNA vaccines encoding IL-2 linked to HPV-16 E7 antigen generate enhanced E7-specific CTL responses and antitumor activity. *Immunol Lett* (2007) **114**(2):86–93. doi:10.1016/j.imlet.2007.09.008
4. Tarpey I, van Loon AA, de Haas N, Davis PJ, Orbell S, Cavanagh D, et al. A recombinant turkey Herpesvirus expressing chicken interleukin-2 increases the protection provided by in ovo vaccination with infectious bursal disease and infectious bronchitis virus. *Vaccine* (2007) **25**(51):8529–35. doi:10.1016/j.vaccine.2007.10.006
5. Davey RT Jr, Murphy RL, Graziano FM, Boswell SL, Pavia AT, Cancio M, et al. Immunologic and virologic effects of subcutaneous interleukin 2 in combination with antiretroviral therapy: a randomized controlled trial. *JAMA* (2000) **284**(2):183–9. doi:10.1001/jama.284.2.183
6. Kovacs JA, Vogel S, Albert JM, Falloon J, Davey RT Jr, Walker RE, et al. Controlled trial of interleukin-2 infusions in patients infected with the human immunodeficiency virus. *N Engl J Med* (1996) **335**(18):1350–6. doi:10.1056/NEJM199610313351803

7. Sereti I, Martinez-Wilson H, Metcalf JA, Baseler MW, Hallahan CW, Hahn B, et al. Long-term effects of intermittent interleukin 2 therapy in patients with HIV infection: characterization of a novel subset of CD4(+)/CD25(+) T cells. *Blood* (2002) **100**(6):2159–67.
8. Natarajan V, Lempicki RA, Sereti I, Badralmaa Y, Adelsberger JW, Metcalf JA, et al. Increased peripheral expansion of naive CD4+ T cells in vivo after IL-2 treatment of patients with HIV infection. *Proc Natl Acad Sci U S A* (2002) **99**(16):10712–7. doi:10.1073/pnas.162352399
9. Ahmadzadeh M, Rosenberg SA. IL-2 administration increases $CD4^+$ CD25(hi) $Foxp3^+$ regulatory T cells in cancer patients. *Blood* (2006) **107**(6):2409–14. doi:10.1182/blood-2005-06-2399
10. Montero E, Alonso L, Perez R, Lage A. Interleukin-2 mastering regulation in cancer and autoimmunity. *Ann N Y Acad Sci* (2007) **1107**:239–50. doi:10.1196/annals.1381.026
11. Tang Q, Adams JY, Penaranda C, Melli K, Piaggio E, Sgouroudis E, et al. Central role of defective interleukin-2 production in the triggering of islet autoimmune destruction. *Immunity* (2008) **28**(5):687–97. doi:10.1016/j.immuni.2008.03.016
12. Grinberg-Bleyer Y, Baeyens A, You S, Elhage R, Fourcade G, Gregoire S, et al. IL-2 reverses established type 1 diabetes in NOD mice by a local effect on pancreatic regulatory T cells. *J Exp Med* (2010) **207**(9):1871–8. doi:10.1084/jem.20100209
13. Setoguchi R, Hori S, Takahashi T, Sakaguchi S. Homeostatic maintenance of natural Foxp3(+) CD25(+) CD4(+) regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *J Exp Med* (2005) **201**(5):723–35. doi:10.1084/jem.20041982
14. Onizuka S, Tawara I, Shimizu J, Sakaguchi S, Fujita T, Nakayama E. Tumor rejection by in vivo administration of anti-CD25 (interleukin-2 receptor alpha) monoclonal antibody. *Cancer Res* (1999) **59**(13):3128–33.
15. Church AC. Clinical advances in therapies targeting the interleukin-2 receptor. *QJM* (2003) **96**(2):91–102. doi:10.1093/qjmed/hcg014
16. Kamimura D, Sawa Y, Sato M, Agung E, Hirano T, Murakami M. IL-2 in vivo activities and antitumor efficacy enhanced by an anti-IL-2 mAb. *J Immunol* (2006) **177**(1):306–14.
17. Mostbock S, Lutsiak ME, Milenic DE, Baidoo K, Schlom J, Sabzevari H. IL-2/anti-IL-2 antibody complex enhances vaccine-mediated antigen-specific CD8(+) T cell responses and increases the ratio of effector/memory CD8(+) T cells to regulatory T cells. *J Immunol* (2008) **180**(7):5118–29.
18. Webster KE, Walters S, Kohler RE, Mrkvan T, Boyman O, Surh CD, et al. In vivo expansion of Treg cells with IL-2-mAb complexes: induction of resistance to EAE and long-term acceptance of isletallografts without immunosuppression. *J Exp Med* (2009) **206**(4):751–60. doi:10.1084/jem.20082824
19. Boyman O, Surh CD, Sprent J. Potential use of IL-2/anti-IL-2 antibody immune complexes for the treatment of cancer and autoimmune disease. *Expert Opin Biol Ther* (2006) **6**(12):1323–31. doi:10.1517/14712598.6.12.1323
20. Boyman O, Kovar M, Rubinstein MP, Surh CD, Sprent J. Selective stimulation of T cell subsets with antibody-cytokine immune complexes. *Science* (2006) **311**:1924–7. doi:10.1126/science.1122927
21. Smith KA. Interleukin-2: inception, impact, and implications. *Science* (1988) **240**(4856):1169–76. doi:10.1126/science.3131876
22. Almeida AR, Legrand N, Papiernik M, Freitas AA. Homeostasis of peripheral $CD4^+$ T cells: IL-2R alpha and IL-2 shape a population of regulatory cells that controls CD4+ T cell numbers. *J Immunol* (2002) **169**(9):4850–60.
23. Grinberg-Bleyer Y, Saadoun D, Baeyens A, Billiard F, Goldstein JD, Grégoire S, et al. Pathogenic T cells have a paradoxical protective effect in murine autoimmune diabetes by boosting Tregs. *J Clin Invest* (2010) **120**(12):4558–68. doi:10.1172/JCI42945
24. Almeida AR, Zaragoza B, Freitas AA. Indexation as a novel mechanism of lymphocyte homeostasis: the number of CD4+CD25+ regulatory T cells is indexed to the number of IL-2-producing cells. *J Immunol* (2006) **177**(1):192–200.
25. Barthlott T, Moncrieffe H, Veldhoen M, Atkins CJ, Christensen J, O'Garra A, et al. CD25+ CD4+ T cells compete with naive CD4+ T cells for IL-2 and exploit it for the induction of IL-10 production. *Int Immunol* (2005) **17**(3):279–88. doi:10.1093/intimm/dxh207
26. Thornton AM, Shevach EM. CD4+CD25+ immunoregulatory T cells suppress polyclonal T cell activation in vitro by inhibiting interleukin 2 production. *J Exp Med* (1998) **188**(2):287–96. doi:10.1084/jem.188.2.287
27. Garcia-Martinez K, Leon K. Modeling the role of IL-2 in the interplay between CD4+ helper and regulatory T cells: studying the impact of IL2 modulation therapies. *Int Immunol* (2012) **24**(7):427–46. doi:10.1093/intimm/dxr120
28. Garcia-Martinez K, Leon K. Modeling the role of IL-2 in the interplay between CD4+ helper and regulatory T cells: assessing general dynamical properties. *J Theor Biol* (2010) **262**(4):720–32. doi:10.1016/j.jtbi.2009.10.025
29. Smith KA. The structure of IL2 bound to the three chains of the IL2 receptor and how signaling occurs. *Med Immunol* (2006) **5**:3. doi:10.1186/1476-9433-5-3
30. Smith KA. The quantal theory of how the immune system discriminates between "self and non-self." *Med Immunol* (2004) **3**(1):3. doi:10.1186/1476-9433-3-3
31. Kuniyasu Y, Takahashi T, Itoh M, Shimizu J, Toda G, Sakaguchi S. Naturally anergic and suppressive CD25(+)CD4(+) T cells as a functionally and phenotypically distinct immunoregulatory T cell subpopulation. *Int Immunol* (2000) **12**(8):1145–55. doi:10.1093/intimm/12.8.1145
32. Létourneau S, Krieg C, Pantaleo G, Boyman O. IL-2- and CD25-dependent immunoregulatory mechanisms in the homeostasis of T-cell subsets. *J Allergy Clin Immunol* (2009) **123**(4):758–62. doi:10.1016/j.jaci.2009.02.011
33. León K, Peréz R, Lage A, Carneiro J. Modelling T-cell-mediated suppression dependent on interactions in multicellular conjugates. *J Theor Biol* (2000) **207**(2):231–54. doi:10.1006/jtbi.2000.2169
34. Carneiro J, Leon K, Caramalho I, van denDoolC, Gardner R, Oliveira V, et al. When three is not a crowd: a crossregulation model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunol Rev* (2007) **216**:48–68.
35. Blattman JN, Grayson JM, Wherry EJ, Kaech SM, Smith KA, Ahmed R. Therapeutic use of IL-2 to enhance antiviral T-cell responses in vivo. *Nat Med* (2003) **9**(5):540–7. doi:10.1038/nm866
36. Cesana GC, DeRaffele G, Cohen S, Moroziewicz D, Mitcham J, Stoutenburg J, et al. Characterization of CD4+CD25+ regulatory T cells in patients treated with high-dose interleukin-2 for metastatic melanoma or renal cell carcinoma. *J Clin Oncol* (2006) **24**(7):1169–77. doi:10.1200/JCO.2005.03.6830
37. Sereti I, Imamichi H, Natarajan V, Imamichi T, Ramchandani MS, Badralmaa Y, et al. In vivo expansion of CD4CD45RO-CD25 T cells expressing foxP3 in IL-2-treated HIV-infected patients. *J Clin Invest* (2005) **115**(7):1839–47. doi:10.1172/JCI24307
38. Rojas G, Pupo A, Leon K, Avellanet J, Carmenate T, Sidhu S. Deciphering the molecular bases of the biological effects of antibodies against Interleukin-2: a versatile platform for fine epitope mapping. *Immunobiology* (2013) **218**(1):105–13. doi:10.1016/j.imbio.2012.02.009
39. Létourneau S, van Leeuwen EM, Krieg C, Martin C, Pantaleo G, Sprent J, et al. IL-2/anti-IL-2 antibody complexes show strong biological activity by avoiding interaction with IL-2 receptor alpha subunit CD25. *Proc Natl Acad Sci U S A* (2010) **107**(5):2171–6. doi:10.1073/pnas.0909384107
40. Carmenate T, Pacios A, Enamorado M, Moreno E, Garcia-Martínez K, Fuente D, et al. Human IL-2 mutein with higher antitumor efficacy than wild type IL-2. *J Immunol* (2013) **190**(12):6230–8. doi:10.4049/jimmunol.1201895
41. Rao BM, Driver I, Lauffenburger DA, Wittrup KD. High-affinity CD25-binding IL-2 mutants potently stimulate persistent T cell growth. *Biochemistry* (2005) **44**(31):10696–701. doi:10.1021/bi050436x
42. Levin AM, Bates DL, Ring AM, Krieg C, Lin JT, Su L, et al. Exploiting a natural conformational switch to engineer an interleukin-2 'superkine.' *Nature* (2012) **484**(7395):529–33. doi:10.1038/nature10975
43. Harvill ET, Fleming JM, Morrison SL. In vivo properties of an IgG3-IL-2 fusion protein. A general strategy for immune potentiation. *J Immunol* (1996) **157**(7):3165–70.
44. Williams MA, Tyznik AJ, Bevan MJ. Interleukin-2 signals during priming are required for secondary expansion of CD8+ memory T cells. *Nature* (2006) **441**(7095):890–3. doi:10.1038/nature04790
45. Kamimura D, Bevan MJ. Naive CD8+ T cells differentiate into protective memory-like cells after IL-2 anti IL-2 complex treatment in vivo. *J Exp Med* (2007) **204**(8):1803–12. doi:10.1084/jem.20070543
46. Zheng SG, Wang J, Wang P, Gray JD, Horwitz DA. IL-2 is essential for TGF-beta to convert naive CD4+CD25− cells to CD25+Foxp3+ regulatory T cells and for expansion of these cells. *J Immunol* (2007) **178**(4):2018–27.
47. Krieg C, Letourneau S, Pantaleo G, Boyman O. Improved IL-2 immunotherapy by selective stimulation of IL-2 receptors on lymphocytes and endothelial cells. *Proc Natl Acad Sci U S A* (2010) **107**:11906. doi:10.1073/pnas.1002569107

48. Wolf M, Schimpl A, Hunig T. Control of T cell hyperactivation in IL-2-deficient mice by CD4(+)CD25(−) and CD4(+)CD25(+) T cells: evidence for two distinct regulatory mechanisms. *Eur J Immunol* (2001) **31**(6): 1637–45. doi:10.1002/1521-4141(200106)31:6<1637::AID-IMMU1637>3.0.CO;2-T
49. Malek TR, Porter BO, Codias EK, Scibelli P, Yu A. Normal lymphoid homeostasis and lack of lethal autoimmunity in mice containing mature T cells with severely impaired IL-2 receptors. *J Immunol* (2000) **164**(6):2905–14.

Conflict of Interest Statement: The authors dclare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 August 2013; paper pending published: 14 September 2013; accepted: 24 November 2013; published online: 11 December 2013.
*Citation: León K, García-Martínez K and Carmenate T (2013) Mathematical models of the impact of IL2 modulation therapies on T cell dynamics. Front. Immunol. **4**:439. doi: 10.3389/fimmu.2013.00439*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 León, García-Martínez and Carmenate. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX A

DYNAMICS IN THE BLOOD COMPARTMENT

Equations for the dynamics in the blood compartment are the following:

$$\frac{d\mathrm{IL2_S}}{dt} = K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2_S^{Ab}} - K_{\mathrm{on}}^{\mathrm{Ab}} \times \frac{1}{N_{\mathrm{A}} \times V_{\mathrm{S}}} \times \mathrm{IL2_S} \times \mathrm{Ab_S} + N_{\mathrm{LN}} \times \left(D_{\mathrm{il}} \frac{\mathrm{IL2}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{il}} \frac{\mathrm{IL2_S}}{V_{\mathrm{S}}} \right) - K_{\mathrm{di}} \times \mathrm{IL2_S} + \Gamma_{\mathrm{i}} \quad \text{(A1)}$$

$$\frac{d\mathrm{IL2m_S}}{dt} = N_{\mathrm{LN}} \times \left(D_{\mathrm{il}} \frac{\mathrm{IL2m}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{il}} \frac{\mathrm{IL2m}_S}{V_{\mathrm{S}}} \right) - K_{\mathrm{di}} \times \mathrm{IL2m_S} + \Gamma_{\mathrm{mi}} \quad \text{(A2)}$$

$$\frac{d\mathrm{Ab_S}}{dt} = K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2_S^{Ab}} - K_{\mathrm{on}}^{\mathrm{Ab}} \times \frac{1}{N_{\mathrm{A}} \times V_{\mathrm{S}}} \times \mathrm{IL2}_S \times \mathrm{Ab_S} + N_{\mathrm{LN}} \times \left(D_{\mathrm{ab}} \frac{\mathrm{Ab}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{ab}} \frac{\mathrm{Ab_S}}{V_{\mathrm{S}}} \right) - K_{\mathrm{da}} \times \mathrm{Ab_S} + \Gamma_{\mathrm{ab}} \quad \text{(A3)}$$

$$\frac{d\mathrm{IL2_S^{Ab}}}{dt} = -K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2_S^{Ab}} + K_{\mathrm{on}}^{\mathrm{Ab}} \times \frac{1}{N_{\mathrm{A}} \times V_{\mathrm{S}}} \times \mathrm{IL2_S} \times \mathrm{Ab_S} + N_{\mathrm{LN}} \times \left(D_{\mathrm{ab}} \frac{\mathrm{IL2^{Ab}}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{ab}} \frac{\mathrm{IL2_S^{Ab}}}{V_{\mathrm{S}}} \right) - K_{\mathrm{da}} \times \mathrm{IL2_S^{Ab}} \quad \text{(A4)}$$

Equations A1–A3 model the dynamics of IL2 ($\mathrm{IL2_s}$), IL2 mutant variants ($\mathrm{IL2m_s}$), and anti-IL2 antibodies ($\mathrm{Ab_s}$) number respectively, while the dynamics of the number of immune-complexes IL2 + anti-IL2 antibodies ($\mathrm{IL2_s^{Ab}}$) is modeled by Eq. A4. The variables and parameters involved in these equations are defined in **Table 1**.

Equations A1, A3, and A4 consider the increase in the number of IL2 and mAbs in the blood due to the dissociation process of immune-complexes with a constant rate ($K_{\mathrm{off}}^{\mathrm{Ab}}$), which corresponds to a decrease in the amount of these complexes (first term in Eqs A1, A3, and A4). The process of formation of immune-complexes, through the association of IL2 and mAb with a constant rate ($K_{\mathrm{on}}^{\mathrm{Ab}}$), is taken into account in the second term in Eqs A1, A3, and A4. The exchange of molecules between blood and peripheral lymph nodes is modeled as a simple diffusion process that balances the molecule concentrations in both compartments (third term in Eqs A1, A3, and A4; first term in Eq. A2). The number of molecules decays exponentially with a constant rate ($K_{\mathrm{di}}, K_{\mathrm{da}}$), due to renal elimination in kidney (fourth term in Eqs A1, A3, and A4 and second term in Eq. A2). Finally, an external source for IL2, IL2m, and mAbs is considered, which causes an increase in the number of these molecules in the compartment (last term in Eqs A1–A3).

APPENDIX B

DYNAMICS OF T CELLS IN THE LYMPH NODE COMPARTMENT

The dynamics of the number of T cells in the lymph node compartment, following the process described above, are modeled with the following set of equations:

$$\frac{d\,E_{\mathrm{N}}}{dt} = \Gamma_{\mathrm{e}} - K_{\mathrm{A}}^{\mathrm{E}} \times E_{\mathrm{N}}^{\mathrm{B}} \times \left(1 - \frac{R_{\mathrm{T}}^{\mathrm{B}}}{s \times A}\right)^{(\mathrm{s}-1)} + \alpha_{\mathrm{E}} \times K_{\mathrm{S}}^{\mathrm{E}} \times \left(1 - \frac{(\mathrm{SigE})^{\mathrm{h}}}{(S_{\mathrm{E}})^{\mathrm{h}} + (\mathrm{SigE})^{\mathrm{h}}}\right) \times E_{\mathrm{A}} + 2\,K_{\mathrm{P}}^{\mathrm{E}} \times E_{\mathrm{C}} - K_{\mathrm{d}}^{\mathrm{E}} \times E_{\mathrm{N}}^{\mathrm{F}} \quad \text{(B1)}$$

$$\frac{d\,E_{\mathrm{A}}}{dt} = K_{\mathrm{A}}^{\mathrm{E}} \times E_{\mathrm{N}}^{\mathrm{B}} \times \left(1 - \frac{R_{\mathrm{T}}^{\mathrm{B}}}{s \times A}\right)^{(\mathrm{s}-1)} - K_{\mathrm{S}}^{\mathrm{E}} \times E_{\mathrm{A}} \quad \text{(B2)}$$

$$\frac{d\,E_{\mathrm{C}}}{dt} = K_{\mathrm{S}}^{\mathrm{E}} \times \left(\frac{(\mathrm{SigE})^{\mathrm{h}}}{(S_{\mathrm{E}})^{\mathrm{h}} + (\mathrm{SigE})^{\mathrm{h}}}\right) \times E_{\mathrm{A}} - K_{\mathrm{P}}^{\mathrm{E}} \times E_{\mathrm{C}} \quad \text{(B3)}$$

$$\frac{d\,R_{\mathrm{N}}}{dt} = \Gamma_{\mathrm{r}} - K_{\mathrm{A}}^{\mathrm{R}} \times R_{\mathrm{N}}^{\mathrm{B}} + \alpha_{\mathrm{R}} \times K_{\mathrm{S}}^{\mathrm{R}} \times \left(1 - \frac{(\mathrm{SigR})^{\mathrm{h}}}{(S_{\mathrm{R}})^{\mathrm{h}} + (\mathrm{SigR})^{\mathrm{h}}}\right) \times R_{\mathrm{A}} + 2\,K_{\mathrm{P}}^{\mathrm{R}} \times R_{\mathrm{C}} - K_{\mathrm{d}}^{\mathrm{R}} \times R_{\mathrm{N}}^{\mathrm{F}} \quad \text{(B4)}$$

$$\frac{d\,R_{\mathrm{A}}}{dt} = K_{\mathrm{A}}^{\mathrm{R}} \times R_{\mathrm{N}}^{\mathrm{B}} - K_{\mathrm{S}}^{\mathrm{R}} \times R_{\mathrm{A}} \quad \text{(B5)}$$

$$\frac{d\,R_{\mathrm{C}}}{dt} = K_{\mathrm{S}}^{\mathrm{R}} \times \left(\frac{(\mathrm{SigR})^{\mathrm{h}}}{(S_{\mathrm{R}})^{\mathrm{h}} + (\mathrm{SigR})^{\mathrm{h}}}\right) \times R_{\mathrm{A}} - K_{\mathrm{P}}^{\mathrm{R}} \times R_{\mathrm{C}} \quad \text{(B6)}$$

$$\frac{d\,M_{\mathrm{A}}}{dt} = -K_{\mathrm{S}}^{\mathrm{M}} \times \left(\frac{(\mathrm{SigM})^{\mathrm{h}}}{(S_{\mathrm{M}})^{\mathrm{h}} + (\mathrm{SigM})^{\mathrm{h}}}\right) \times M_{\mathrm{A}} + 2 \times K_{\mathrm{P}}^{\mathrm{M}} \times M_{\mathrm{C}} - K_{\mathrm{d}}^{\mathrm{M}} \times M_{\mathrm{A}} \quad \text{(B7)}$$

$$\frac{d\,M_{\mathrm{C}}}{dt} = K_{\mathrm{S}}^{\mathrm{M}} \times \left(\frac{(\mathrm{SigM})^{\mathrm{h}}}{(S_{\mathrm{M}})^{\mathrm{h}} + (\mathrm{SigM})^{\mathrm{h}}}\right) \times M_{\mathrm{A}} - K_{\mathrm{P}}^{\mathrm{M}} \times M_{\mathrm{C}} \quad \text{(B8)}$$

$$\frac{K^{\mathrm{E}}}{V_{\mathrm{LN}}} = E_{\mathrm{l}}^{\mathrm{B}}/(E_{\mathrm{l}}^{\mathrm{F}} \times F);\ \frac{K^{\mathrm{R}}}{V_{\mathrm{LN}}} = R_{\mathrm{l}}^{\mathrm{B}}/(R_{\mathrm{l}}^{\mathrm{F}} \times F);\ \frac{K^{\mathrm{M}}}{V_{\mathrm{LN}}} = M_{\mathrm{l}}^{\mathrm{B}}/(M_{\mathrm{l}}^{\mathrm{F}} \times F) \quad \text{(B9)}$$

$$F = s \times A - \sum_{l} E_{\mathrm{l}}^{\mathrm{B}} - \sum_{l} R_{\mathrm{l}}^{\mathrm{B}} - \sum_{l} M_{\mathrm{l}}^{\mathrm{B}};\quad \forall l \in \{\mathrm{N}, \mathrm{A}, \mathrm{C}\} \quad \text{(B10)}$$

The dynamics of the number of Helper and Regulatory CD4$^+$ T cells, on their three different functional states of their life cycle [resting (E_N, R_N), activated (E_A, R_A) and cycling (E_C, R_C) cells], is modeled using Eqs B1–B6, respectively; while the dynamics of the number of Memory CD8$^+$ T cells on its two functional states [activated (M_A) and cycling (M_C) cells] is modeled with Eqs B7 and B8. The process of conjugation of T cells, on their different functional states, with their cognate APCs is modeled assuming quasi-steady state equilibrium, which leads to the equations presented in Eq. B9 [see a more detailed explanation of the derivation of these equations in (28)]. In Eq. B9, the symbolic label l denotes the functional state of the cell ($l = N$: resting, $l = A$: activated, $l = C$: cycling); and the superscript B and F denotes the cells conjugated to APC or free, respectively. The definitions of variables and parameters in Eqs B1–B9 are resumed in **Table 2**.

Equations B1–B6 considered that resting E and R cells are produced by the thymus (first term in Eqs B1 and B4), and they die with a constant rate K_d^E and K_d^R, respectively (last term in Eqs B1 and B4). Resting cells become activated after conjugation with APCs, process which is inhibited in E cells by the presence of co-conjugated R cells in the same APC (second term in Eqs B1 and B4; first term in Eqs B2 and B5). Activated T cells require enough cytokine derived signals to become cycling cells (first term in Eqs B3 and B6). The fraction of activated cells obtaining these signals is computed with a sigmoid function of the mean number of bound cytokines signaling receptors per cell (SigE, SigR). In the absence of these signals, a fraction α of the activated cells revert to the resting state (third term in Eqs B1 and B4) and the remaining fraction $(1-\alpha)$ simply die. The cycling E and R cells divide producing two new resting cells with a constant rate K_p^E and K_p^R, respectively (fourth term in Eqs B1 and B4; second term in Eqs B3 and B6).

The Eqs B7 and B8 describe the dynamics of memory CD8$^+$ T cells. In these equations, is modeled the dynamics of M cells analogous that for E cells. The only difference is that M cells are considered pre-activated cells, which become cycling in response to cytokine signals (first term in Eqs B7 and B8). The cycling M cells divide producing two new activated cells with a constant rate K_p^M (second term in Eqs B7 and B8). The activated cells die with a constant rate K_d^M (last term in Eq. B7).

APPENDIX C

DYNAMICS OF MOLECULES IN THE LYMPH NODE

The equations in the model describing the dynamics of the number of molecules circulating in the Lymph Node (IL2, anti-IL2 antibodies, and immune-complexes) and the number of complexes IL2-IL2R and IL2-mAb-IL2R formed in a single cell membrane are the following:

$$\frac{d\mathrm{IL2}}{dt} = K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2}^{\mathrm{Ab}} - \frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^{\mathrm{Ab}} \times \mathrm{IL2} \times \mathrm{Ab} - \left(D_{\mathrm{il}} \frac{\mathrm{IL2}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{il}} \frac{\mathrm{IL2_S}}{V_S} \right) + K_{\mathrm{pi}} \times K_A^E \times E_N^B \times \left(1 - \frac{R_T^B}{s \times A} \right)^{(s-1)} + \sum_j K_{\mathrm{off}}^j \times \left[\sum_l (C_j^{E_l} \times E_l) + \sum_l (C_j^{R_l} \times R_l) + \sum_l (C_j^{M_l} \times M_l) \right] - \sum_j \frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^j \times \mathrm{IL2} \times \left[\sum_l (P_j^{E_l} \times E_l) + \sum_l (P_j^{R_l} \times R_l) + \sum_l (P_j^{M_l} \times M_l) \right] \quad \text{(C1)}$$

$$\frac{d\mathrm{IL2m}}{dt} = -\left(D_{\mathrm{il}} \frac{\mathrm{IL2m}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{il}} \frac{\mathrm{IL2m_S}}{V_S} \right) + \sum_j K_{\mathrm{off}}^j \times \left[\sum_l (\mathrm{Cm}_j^{E_l} \times E_l) + \sum_l (\mathrm{Cm}_j^{R_l} \times R_l) + \sum_l (\mathrm{Cm}_j^{M_l} \times M_l) \right] - \sum_j f_j \times \frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^j \times \mathrm{IL2m} \times \left[\sum_l (P_j^{E_l} \times E_l) + \sum_l (P_j^{R_l} \times R_l) + \sum_l (P_j^{M_l} \times M_l) \right] \quad \text{(C2)}$$

$$\frac{d\mathrm{Ab}}{dt} = K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2}^{\mathrm{Ab}} - \frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^{\mathrm{Ab}} \times \mathrm{IL2} \times \mathrm{Ab} - \left(D_{\mathrm{ab}} \frac{\mathrm{Ab}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{ab}} \frac{\mathrm{Ab_S}}{V_S} \right) + \sum_j (1 - N_j) \times \left[K_{\mathrm{off}}^{\mathrm{Ab}} \times \sum_l (\mathrm{CAb}_j^{E_l} \times E_l + \mathrm{CAb}_j^{R_l} \times R_l + \mathrm{CAb}_j^{M_l} \times M_l) \right] - \sum_j (1 - N_j) \times \left[\frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^{\mathrm{Ab}} \times \mathrm{Ab} \times \sum_l \left(C_j^{E_l} \times E_l + C_j^{R_l} \times R_l + C_j^{M_l} \times M_l \right) \right] \quad \text{(C3)}$$

$$\frac{d\mathrm{IL2^{Ab}}}{dt} = -K_{\mathrm{off}}^{\mathrm{Ab}} \times \mathrm{IL2}^{\mathrm{Ab}} + \frac{1}{N_A \times \mathrm{fve} \times V_{\mathrm{LN}}} K_{\mathrm{on}}^{\mathrm{Ab}} \times \mathrm{IL2} \times \mathrm{Ab} - \left(D_{\mathrm{ab}} \frac{\mathrm{IL2^{Ab}}}{\mathrm{fve} \times V_{\mathrm{LN}}} - D_{\mathrm{ab}} \frac{\mathrm{IL2_S^{Ab}}}{V_S} \right)$$

$$+\sum_j (1-N_j)\times K_{\text{off}}^{\text{j}} \times \left[\sum_l (\text{CAb}_{\text{j}}^{\text{E}_l}\times E_l)+\sum_l (\text{CAb}_{\text{j}}^{\text{R}_l}\times R_l) + \sum_l (\text{CAb}_{\text{j}}^{\text{M}_l}\times M_l)\right] - \sum_j (1-N_j) \times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} K_{\text{on}}^{\text{j}}\times \text{IL2}^{\text{Ab}} \times \left[\sum_l (P_{\text{j}}^{\text{E}_l}\times E_l)+\sum_l (P_{\text{j}}^{\text{R}_l}\times R_l) + \sum_l (P_{\text{j}}^{\text{M}_l}\times M_l)\right] \quad \text{(C4)}$$

$$\frac{d\text{Cm}_\alpha^{\text{E}_l}}{dt} = f_\alpha \times K_{\text{on}}^{\alpha} \times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} \times \text{IL2m}\times P_\alpha^{\text{E}_l} - K_{\text{off}}^{\alpha}\times \text{Cm}_\alpha^{\text{E}_l} - f_\beta \times K_{\text{on}}^{\alpha\beta} \times \text{Cm}_\alpha^{\text{E}_l}\times P_\beta^{\text{E}_l} + K_{\text{off}}^{\alpha\beta}\times \text{Tm}^{\text{E}_l} \quad \text{(C5)}$$

$$\frac{d\text{Cm}_\beta^{\text{E}_l}}{dt} = f_\beta \times K_{\text{on}}^{\beta} \times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} \times \text{IL2m}\times P_\beta^{\text{E}_l} - K_{\text{off}}^{\beta}\times \text{Cm}_\beta^{\text{E}_l} - f_\alpha \times K_{\text{on}}^{\beta\alpha} \times \text{Cm}_\beta^{\text{E}_l}\times P_\alpha^{\text{E}_l} + K_{\text{off}}^{\beta\alpha}\times \text{Tm}^{\text{E}_l} - K_{\text{in}}\times \text{Cm}_\beta^{\text{E}_l} \quad \text{(C6)}$$

$$\frac{d\text{Tm}^{\text{E}_l}}{dt} = f_\beta \times K_{\text{on}}^{\alpha\beta}\times \text{Cm}_\alpha^{\text{E}_l}\times P_\beta^{\text{E}_l} - K_{\text{off}}^{\alpha\beta}\times \text{Tm}^{\text{E}_l} + f_\alpha \times K_{\text{on}}^{\beta\alpha}\times \text{Cm}_\beta^{\text{E}_l}\times P_\alpha^{\text{E}_l} - K_{\text{off}}^{\beta\alpha}\times \text{Tm}^{\text{E}_l} - K_{\text{in}}\times \text{Tm}^{\text{E}_l} \quad \text{(C7)}$$

$$\frac{d\text{C}_\alpha^{\text{E}_l}}{dt} = K_{\text{on}}^{\alpha}\times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}}\times \text{IL2}\times P_\alpha^{\text{E}_l} - K_{\text{off}}^{\alpha} \times C_\alpha^{\text{E}_l} - K_{\text{on}}^{\alpha\beta}\times C_\alpha^{\text{E}_l}\times P_\beta^{\text{E}_l} + K_{\text{off}}^{\alpha\beta}\times T^{\text{E}_l} + (1-N_\alpha)\times\left(-\frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} K_{\text{on}}^{\text{Ab}} \times C_\alpha^{\text{E}_l}\times \text{Ab} + K_{\text{off}}^{\text{Ab}}\times \text{CAb}_\alpha^{\text{E}_l}\right) \quad \text{(C8)}$$

$$\frac{dC_\beta^{E_l}}{dt} = K_{on}^{\beta}\times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}}\times \text{IL2}\times P_\beta^{\text{E}_l} - K_{\text{off}}^{\beta} \times C_\beta^{\text{E}_l} - K_{\text{on}}^{\beta\alpha}\times C_\beta^{\text{E}_l}\times P_\alpha^{\text{E}_l} + K_{\text{off}}^{\beta\alpha}\times T^{\text{E}_l} + (1-N_\beta)\times\left(-\frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} K_{\text{on}}^{\text{Ab}} \times C_\beta^{\text{E}_l}\times \text{Ab} + K_{\text{off}}^{\text{Ab}}\times \text{CAb}_\beta^{\text{E}_l}\right) - K_{\text{in}}\times C_\beta^{\text{E}_l} \quad \text{(C9)}$$

$$\frac{dT^{\text{E}_l}}{dt} = K_{\text{on}}^{\alpha\beta}\times C_\alpha^{\text{E}_l}\times P_\beta^{\text{E}_l} - K_{\text{off}}^{\alpha\beta}\times T^{\text{E}_l} + K_{\text{on}}^{\beta\alpha}\times C_\beta^{\text{E}_l} \times P_\alpha^{\text{E}_l} - K_{\text{off}}^{\beta\alpha}\times T^{\text{E}_l} - K_{\text{in}}\times T^{\text{E}_l} \quad \text{(C10)}$$

$$\frac{d\text{CAb}_\alpha^{\text{E}_l}}{dt} = (1-N_\alpha)\times\left(K_{\text{on}}^{\alpha}\times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}}\times \text{IL2}^{\text{Ab}} \times P_\alpha^{\text{E}_l} - K_{\text{off}}^{\alpha}\times \text{CAb}_\alpha^{\text{E}_l}\right) + (1-N_\alpha) \times\left(\frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} K_{\text{on}}^{\text{Ab}}\times C_\alpha^{\text{E}_l}\times \text{Ab} - K_{\text{off}}^{\text{Ab}}\times \text{CAb}_\alpha^{\text{E}_l}\right) \quad \text{(C11)}$$

$$\frac{d\text{CAb}_\beta^{\text{E}_l}}{dt} = (1-N_\beta)\times\left(K_{\text{on}}^{\beta}\times \frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} \times\text{IL2}^{\text{Ab}}\times P_\beta^{\text{E}_l} - K_{\text{off}}^{\beta}\times \text{CAb}_\beta^{\text{E}_l}\right) + (1-N_\beta) \times\left(\frac{1}{N_{\text{A}}\times \text{fve}\times V_{\text{LN}}} K_{\text{on}}^{\text{Ab}}\times C_\beta^{\text{E}_l}\times \text{Ab} - K_{\text{off}}^{\text{Ab}}\times \text{CAb}_\beta^{\text{E}_l}\right) - K_{\text{in}}\times \text{CAb}_\beta^{\text{E}_l} \quad \text{(C12)}$$

$$P_\alpha^{\text{E}_l} = \text{Ra}^{\text{E}_l} - C_\alpha^{\text{E}_l} - \text{CAb}_\alpha^{\text{E}_l} - T^{\text{E}_l} - \text{Cm}_\alpha^{\text{E}_l} - \text{Tm}^{\text{E}_l},$$
$$P_\beta^{\text{E}_l} = \text{Rb}^{\text{E}_l} - C_\beta^{\text{E}_l} - \text{CAb}_\beta^{\text{E}_l} - T^{\text{E}_l} - \text{Cm}_\beta^{\text{E}_l} - \text{Tm}^{\text{E}_l} \quad \text{(C13)}$$

$$\text{SigE} = C_\beta^{E_A} + T^{E_A} + \text{Cm}_\beta^{E_A} + \text{Tm}^{E_A} + \text{CAb}_\beta^{E_A} + \text{il}\alpha_{E_A},$$
$$\text{SigR} = C_\beta^{R_A} + T^{R_A} + \text{Cm}_\beta^{R_A} + \text{Tm}^{R_A} + \text{CAb}_\beta^{R_A}, \quad \text{(C14)}$$
$$\text{SigM} = C_\beta^{M_A} + T^{M_A} + \text{Cm}_\beta^{M_A} + \text{Tm}^{M_A} + \text{CAb}_\beta^{M_A} + \text{il}\alpha_{M_A}$$

The dynamics of the number of IL2 (IL2), IL2 mutants (IL2m), mAbs (Ab), and immune-complexes (IL2^{Ab}) in the lymph node is modeled using Eqs C1 and C4; while the dynamics of the number of IL2-IL2R complexes (C_α^{E}, C_β^{E}, T^{E}); IL2m-IL2R complexes ($\text{Cm}_\alpha^{\text{E}}$, $\text{Cm}_\beta^{\text{E}}$, Tm^{E}); and IL2-IL2R-mAbs complexes $\text{CAb}_\alpha^{\text{E}}, \text{CAb}_\beta^{\text{E}}$ per cell are modeled following equations (22–24); (19–21); and (25, 26), respectively. Note that, to simplify, we only present here the equations corresponding to the IL2 and IL2m complexes formed at the E cell membrane. Equivalent equations are written for R and M cells. Algebraic relations are provided in Eqs C13 and C14, for the amount of free alpha (P_α^{E}) and beta chains (P_β^{E}) of the IL2 receptor per E cell, and the mean number of bound cytokines signaling receptors per activated E (SigE), R (SigR), and M (SigM) cell. Note that, the terms SigE, SigR, and SigM in Eq. C14

are the one used in the equations for the dynamics of T cells, related with the process of cells receiving cytokine derived signals from IL2 or alternative cytokines (see section 1.2). The variables and parameters used in Ref. (15–28) are defined in **Tables 1–3**.

In Eqs C1, C3, and C4, the first and second terms correspond to the processes of dissociation and formation of immune-complexes. Additionally, due to the presence of T cells in the Lymph Node, we consider in these equations the dissociation and association processes of IL2 and IL2-mAb complexes with the alpha or beta IL2R chains in the cell membranes. In this sense, is taken into account the increase in the amount of free IL2, IL2m, and immune-complexes, due to the dissociation process of IL2-IL2R, IL2m-IL2R, and IL2-mAb-IL2R complexes respectively (third terms in Eqs C1 and C4 and second term in Eq. C2). On the other hand, the association of free IL2, IL2m, and immune-complexes to free alpha or beta IL2R chains is considered to increase the amount of IL2-IL2R and IL2-mAb-IL2R complexes (fourth terms in Eqs C1 and C4, third term in Eq. C2). Additionally, is modeled the processes where mAbs can be dissociated from IL2-mAb-IL2R complexes, increasing the number of free mAbs (third terms in Eq. C3); and the process where free mAbs associate to IL2-IL2R complexes in the cell membrane (fourth terms in Eq. C3). The symbolic labels l and j, appearing in the equations (15–18), denote respectively the functional state of the cell ($l = N$: resting, $l = A$: activated, $l = C$: cycling) and the different IL2R chains ($j = \alpha$ alpha chain and $j = \beta$ beta dimer chain). Finally, the production of IL2 endogenous by activated E cells, which can be inhibited during cell activation by the presence of R cells co-conjugated in the same APC, is considered to increase the amount of this cytokine in the Lymph Node (last term in Eq. C1). The properties of different IL2m is controlled in the model by the parameters f_α and f_β which multiply the association of constant of this molecules to the alpha and beta chain of the IL2 receptor.

The formation of high affinity IL2-IL2R and IL2m-IL2R complexes in a cell membrane is modeled as a two-step process, using equations (22–24) and (19–21) respectively. Firstly, free IL2 or IL2m binds to the available free alpha or beta chains of the IL2R, forming the intermediate or low affinity IL2-IL2R complexes respectively (first terms in Eqs C8, C9 and C5, C6), as mentioned above for the dynamics of IL2. By the corresponding dissociation process are recovered free molecules and receptor chains (second term in Eqs C8, C9 and C5, C6). The association process of intermediate or low affinity IL2-IL2R complexes with the remaining IL2 receptor chain, leads to the formation of high affinity IL2-IL2R complexes (third term in Eqs C8, C9 and C5, C6), and first and third terms in Eqs C10 and C7. The dissociation of these complexes is modeled in the fourth term in Eqs C9, C10 and C5, C6 and second and fourth terms in Eqs C10 and C7. The internalization of IL2 and IL2m forming complexes with IL2Rs is modeled considering that it only occurs for signaling IL2-IL2R complexes requiring binding to the beta chain (last term in Eqs C9, C10 and C6, C7).

The formation of IL2-mAb-IL2R complexes in the cell membrane is modeled in Eqs C11 and C12, and in the fifth term in Eqs C8 and C9. In this sense, we consider the association and dissociation processes of free immune-complexes with the alpha or beta IL2R chains (first term in Eqs C11 and C12); and the same processes for free mAbs with IL2-IL2R complexes in the cell membrane (fifth term in Eqs C8 and C9); second term in Eqs C11 and C12). The possibility of formation of intermediate or low affinity IL2-mAb-IL2R complexes depend on the IL2 interface that mAbs recognize (controlled in simulations by the parameter N_j, see **Table 3**). We don't consider the formation of high affinity IL2-mAb-IL2R complexes, due to association of antibodies with the high affinity IL2-IL2R complexes or the association of intermediate or low affinity IL2-mAb-IL2R complexes with the remaining IL2 receptor chain, because we are studying mAbs that bind to the alpha or beta interface of the IL2 which will block the formation of these complexes. Finally, the internalization of IL2 as immune-complexes bound to the beta chains of IL2Rs in the cell membrane is also modeled (last term in Eq. C12).

frontiers in
IMMUNOLOGY

MINI REVIEW ARTICLE
published: 18 November 2013
doi: 10.3389/fimmu.2013.00378

Mechanisms underlying CD4+ Treg immune regulation in the adult: from experiments to models

Marta Caridade[1,2], Luis Graca[1,2]*‡ and Ruy M. Ribeiro[3]*†‡

[1] *Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisbon, Portugal*
[2] *Instituto Gulbenkian de Ciência, Oeiras, Portugal*
[3] *Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Akihiko Yoshimura, Keio University, Japan
Fernando A. Arosa, University of Beira Interior, Portugal

****Correspondence:***
Luis Graca, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Avenida Prof. Egas Moniz, 1649-028 Lisbon, Portugal
e-mail: lgraca@fm.ul.pt;
Ruy M. Ribeiro, Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
e-mail: ruy@lanl.gov

†Present address:
Ruy M. Ribeiro, Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisbon, Portugal

‡Luis Graca and Ruy M. Ribeiro are joint senior authors.

To maintain immunological balance the organism has to be tolerant to self while remaining competent to mount an effective immune response against third-party antigens. An important mechanism of this immune regulation involves the action of regulatory T-cell (Tregs). In this mini-review, we discuss some of the known and proposed mechanisms by which Tregs exert their influence in the context of immune regulation, and the contribution of mathematical modeling for these mechanistic studies. These models explore the mechanisms of action of regulatory T cells, and include hypotheses of multiple signals, delivered through simultaneous antigen-presenting cell (APC) conjugation; interaction of feedback loops between APC, Tregs, and effector cells; or production of specific cytokines that act on effector cells. As the field matures, and competing models are winnowed out, it is likely that we will be able to quantify how tolerance-inducing strategies, such as CD4-blockade, affect T-cell dynamics and what mechanisms explain the observed behavior of T-cell based tolerance.

Keywords: Tregs, mathematical models, CD4-blockade, regulation, tolerance

INTRODUCTION

Immunological tolerance can be defined as the state of unresponsiveness to an antigen, following prior contact with that antigen, where the host remains competent to mount an effective immune response against third-party antigens. Accomplishing therapeutic induced tolerance has been one of the major goals of immunology ever since the pioneering work of Medawar and colleagues (1).

There is a need to keep a balance between aggressive cells and cells that maintain tolerance to self. On occasions this balance can be disrupted originating either autoimmunity, when mechanisms leading to self-tolerance fail, or immunodeficiency and susceptibility to infection when the immune system is not able to mount a proper immune response. Usually, however, the immune system shows a significant capacity for self-tolerance, in spite of its equally efficient performance in the protection from foreign microbes. The ability to orchestrate protective immune responses is also the major hurdle impeding successful transplantation therapies and hinders the efficacy of therapeutic administration of foreign proteins and genes.

Random rearrangement of T-cell receptors (TCR) during cellular maturation leads to T cells that will recognize self-antigens. It was a long held assumption that central tolerance, by means of negative selection of autoreactive lymphocyte clones, could on its own account for the establishment of self-tolerance. Without such a censoring mechanism these autoreactive cells could eventually lead to autoimmune disease. Indeed thymocytes must survive the process of negative selection, which eliminates cells whose TCRs bind too avidly to self-antigens (2–4). The apoptosis of these thymocytes will prevent migration of autoreactive T cells to the periphery and prevent autoimmunity. Conversely, absence of negative selecting self-peptide-MHC complexes in the thymic medulla leads to an increase in mature autoreactive T cells (5, 6).

However, not all self-antigens are presented in the thymus, and some developing autoreactive T cells never encounter their antigens, eventually migrating to the periphery. Thus, although central tolerance contributes to the deletion of a large number of potentially autoreactive T cells, some autoreactive clones can be found in the periphery of healthy individuals (7, 8). There are, therefore, mechanisms that operate in the periphery (i.e., outside the thymus) to establish self-tolerance toward autoreactive clones that escape thymic negative selection.

Initially, one such mechanism was thought to be mediated by T-cell anergy, described as the functional state in which T cells remain viable but unable to respond to optimal stimulation through the TCR and co-stimulatory ligands (9), i.e., unable to proliferate or to produce interleukin-2 (IL-2) (10, 11). The first observation of anergy was made with purified human $CD4^+$ T cells stimulated with large quantities of peptide antigens (10). It was noted that after antigen stimulation there was down-regulation of TCR and this was associated with the molecular mechanism for anergy (12). Subsequent studies with mouse $CD4^+$ T cells suggested that occupancy of the TCR without any other signals was responsible for the induction of this state (13, 14).

Interestingly, anergic T cells were capable of suppressing proliferation of naïve T cells *in vitro* (15) and *in vivo* (16). In addition, anergic T cells have been shown to inhibit the antigen-presenting function and survival of dendritic cells (17). These and other observations led to the proposal of the "civil service model" (18), postulating that antigen-specific unresponsive cells can interfere with the generation of help by co-localizing with other T cells and competing for elements in the microenvironment (such as adhesion molecules or cytokines).

However, it was not clear how T cells would become anergic *in vivo*, and whether such mechanism was enough to maintain tolerance. More recently, a specific T cell subset, termed regulatory T (Treg) cells, gained prominence as being a key mechanism maintaining peripheral self-tolerance (19, 20). With hindsight, it is likely that many of the features of anergic T cells are a consequence of Treg function.

REGULATORY T CELLS

In 1995 Sakaguchi et al. (19) showed that depletion of a minor population of $CD4^+$ T cells constitutively expressing CD25 [IL-2 receptor α-chain (IL-2Rα)] led to the generation of a spectrum of autoimmune diseases when transferred to immune-compromised recipients. In addition, the co-transfer of $CD25^+$ T cells prevented the pathology.

Based on this CD25 marker, a population of natural (thymus-derived) regulatory T cells was identified in the resting immune system, both in mice and in humans (21). Subsequent studies showed that these cells express forkhead box transcription factor 3 (Foxp3) and this finding led to the definite establishment of a Treg subset (22–24). There is now abundant evidence that these regulatory T cells are actively engaged in the maintenance of self-tolerance (25). Furthermore, depletion of $Foxp3^+$ Tregs originates fatal multi-organ autoimmunity. The phenotype of this disease is virtually indistinguishable from the IPEX syndrome, caused by Foxp3 mutations in humans and equivalent to the Scurfy phenotype in mice (26–28).

THYMIC TREG CELLS

The Treg cells that develop in the thymus, first described as naturally occurring regulatory T cells (nTregs) appear to be selected for self-antigen/MHC expressed by thymic epithelial cells (29, 30), in a process that requires TCR triggering in the presence of co-stimulation (31, 32), but dispenses TGF-β and IL-2 (33, 34). Early studies with Treg cells showed that these cells express CD25, CD5, and cytotoxic T lymphocyte antigen 4 (CTLA-4), which are all induced upon TCR stimulation (19).

In the periphery, nTregs represent around 6–10% of the overall $CD4^+$ T-cell population. In order to be sustained they need continuous TCR triggering and co-stimulation in the presence of IL-2 (35–37), making IL-2 essential for natural Treg pool maintenance in the periphery (38). Comparative analysis of polyclonal TCR repertoires showed that TCR sequences from Treg cells were of broader variety and only partially overlapping with the ones from non-Treg cells (39). Some studies have shown that antigen-specific Treg cells are more potent at suppressing the induction of autoimmune disease than polyclonal populations (40). However, other studies have also shown that polyclonal Tregs are able to suppress independently of their specificity (41). Thus, Tregs with one antigen-specificity can suppress effector cells with many other antigen-specificities by bystander suppression. Moreover, transplantation studies have shown that Tregs can display a phenomenon called "linked suppression," where they can be activated in an antigen-specific manner, and subsequently suppress responses to unrelated antigens presented by the same cells (42). Tregs show a third property called infectious tolerance by which one population of Treg cells creates a regulatory milieu that promotes the outgrowth of a new population of Treg cells with antigen-specificities distinct from those of the original population, as long as the new antigen is present in the same tissue as the antigen recognized by the original Treg cell (43–45).

PERIPHERAL TREG INDUCTION

Besides nTreg, of thymic origin, it has become apparent that induced regulatory T cells (iTreg) also exist in the periphery (46, 47). After the discovery of the key role for Foxp3 in Tregs, it was demonstrated that it was possible for non-Treg cells to acquire both Foxp3 and the regulatory functions associated with it, therefore becoming Treg cells themselves (46, 48, 49).

It is likely that peripheral induction of iTreg occurs in response to non-self antigens like food, allergens, and commensal bacteria (39). Early evidence for *in vivo* peripheral conversion was derived from adoptive cell transfer experiments in which polyclonal $CD4^+$ $CD25^-$ naïve T cells were injected into lymphopenic mice or mice containing a monoclonal T cell repertoire devoid of nTregs, or when tolerance was imposed on monoclonal populations without Treg cells (49–51). In these conditions, homeostatic proliferation of the donor cells could be observed and part of the donor cell population became $CD25^+CTLA\text{-}4^+GITR^+Foxp3^+$ and acquired suppressive activity. Additionally, when congenitally marked $CD4^+$ $CD25^-$ T cells were transferred to WT hosts, 10% of those converted into $CD4^+$ $CD25^+$ $Foxp3^+$ T cells, within 6 weeks (52).

It was first shown *in vitro* that TCR activation in the presence of TGF-β would lead to Treg conversion (53). Subsequent studies supported this observation and demonstrated that iTreg conversion could be greatly enhanced by suboptimal TCR signals or a combination of strong TCR signals with high doses of TGF-β (47, 53–57). *In vivo* it is possible to induce oral tolerance by giving the antigen in the drinking water (58), or to induce transplantation tolerance using non-depleting anti-CD4 at the time of transplantation (48, 59). In both cases, tolerance induction is accompanied

by induction of Foxp3$^+$ cells, in a process that requires TGF-β. In addition to these, many other factors influence the induction of Tregs both *in vitro* and *in vivo*, such as the co-stimulation environment, the strength of TCR signaling, mTOR inhibition with rapamycin, and low levels of essential amino-acids (44, 57, 60–69).

MECHANISMS OF ACTION OF TREG CELLS

In spite of intensive study of Tregs and their properties, the specific mechanisms by which they control immune responses are still not fully understood. There are several proposed mechanisms with experimental support, but it is likely that no single mechanism is responsible for the full range of biological phenomena involving Tregs (70). And it is also likely that in different milieu distinct mechanisms and even alternative subsets of regulatory cells are involved in tuning the immune response (71).

In **Figure 1**, we summarize five putative mechanisms of Treg function: (i) modulation of antigen-presenting cell (APC) activity through Treg engagement of co-stimulatory receptors on the surface of APC, leading to weak or abrogated signals from APC to naive/effector cells; (ii) Treg secretion of cytokines, such as IL10 and TGFβ, suppressing the activity of effector cells and APC; (iii) under certain circumstances, Tregs could have a direct cytotoxic effect, through the production of perforin/granzyme and induction of apoptosis in effector cells; (iv) Tregs may also cause metabolic disruption, for example stimulating APCs to produce enzymes that consume essential amino-acids, preventing naive/effector cell proliferation, and in the presence of TGFβ may induce the expression of Foxp3 in naive cells (i.e., they become Tregs); (v) Tregs could also compete with effectors cells for APC signals or cytokines, such as IL2.

There is mounting evidence [reviewed in (72)] that Treg cells exert their effects on different cell types, including CD4$^+$ and CD8$^+$ T cells, B cells, natural killer T cells (NKT), and DCs (70). The action of Tregs can be mediated by secretion of immunosuppressive cytokines, such as IL-10, TGF-β, IL-35, and galectin-1 (72) or by cell-dependent mechanisms through molecules such as GITR, CTLA-4, CD39, CD73, and LAG-3 (70). The spectrum of effect of Tregs on their targets goes from modifying the functional properties of other immune cells, such as down-regulating transcription of IL-2 (70, 71, 73), and other important growth factors; to actually killing those cells through granzyme B and perforin (70, 73–77). For example, there is evidence that Tregs can kill both immature and mature DCs (74).

Furthermore, Tregs may convert APCs to become themselves immunosuppressive (78). It has also been proposed that Tregs act by competing with other cells for growth factors, particularly IL-2 (79, 80). One possible outcome of these interactions is that other cells become themselves Foxp3$^+$ regulatory cells (45).

These and other suppressive mechanisms may be operational depending on the microenvironment, biological context, and

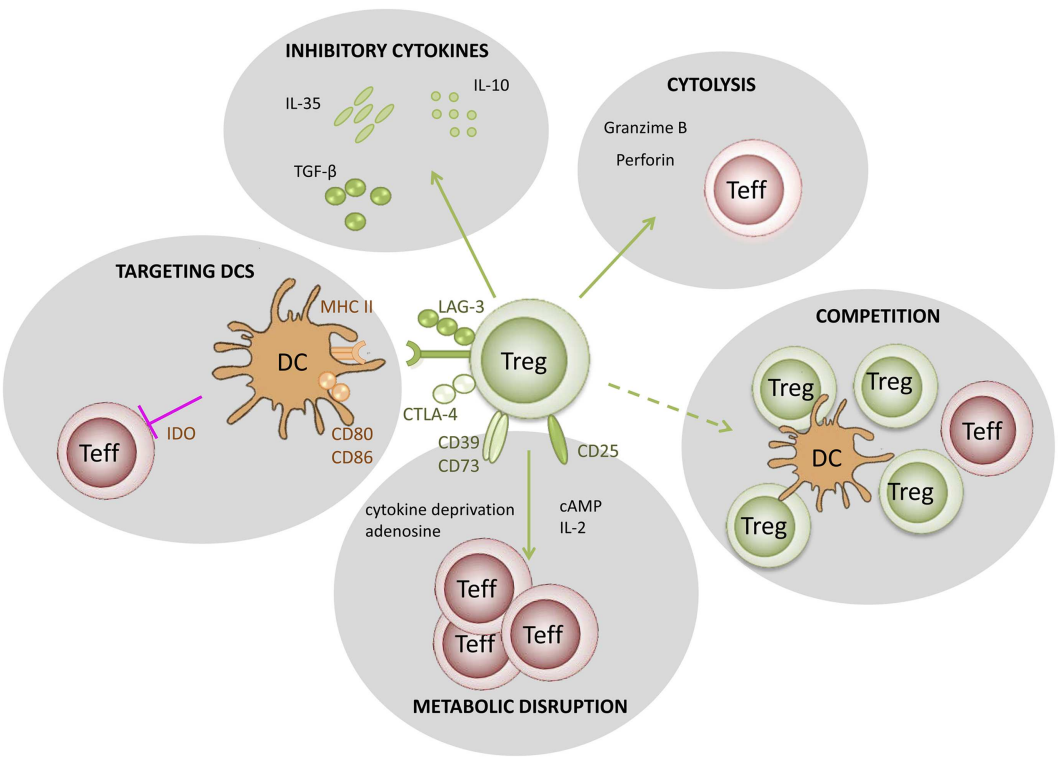


FIGURE 1 | Putative mechanisms used by regulatory T cells. (1) Targeting DCs – modulation of antigen-presenting cell activity through Treg engagement of co-stimulatory receptors on the DC surface, leading to weak or abrogated signals to naïve/effector T cells; (2) Metabolic disruption – includes cytokine deprivation, cyclic AMP-mediated inhibition, and adenosine receptor (A2A)-mediated immunosuppression; (3) Competition – for critical cytokines, such as IL-2, or direct disruption of effector cell engagement with APCs; (4) Cytolysis – direct cytotoxic effect through the production of Granzyme B and Perforin and consequent apoptosis of effector T cells or APCs; (5) Production of inhibitory cytokines – including IL-10, IL-35, and TGF-β.

immune response. For instance, IL-10 producing cells are more abundant in lamina propria (81, 82) and perforin or granzyme expressing Tregs are predominant in tumor environments (83).

MATHEMATICAL MODELING

Due to the complexity of the mechanisms and interactions involved in the processes of immune tolerance, mathematical modeling has been used as a tool to explore different conceptual frameworks of immunological tolerance. Many studies have analyzed the dynamics of thymocyte development, with positive and negative selection, as a mechanism of central tolerance (84–91).

Many other studies have focused on modeling the putative mechanisms of Treg suppression in the periphery. In these models, typically the dynamics of Tregs, effector cells, and APCs are studied to find the interaction mechanisms in the model that qualitatively reflect the experimental knowledge. For the purposes of this review, we can divide the models proposed in three categories, although there is overlap between these in some studies: (i) models that analyze different putative mechanisms of action of Tregs (**Table 1**); (ii) models that analyze the effects of Tregs on different processes, such as the immune response to pathogens and tumors, or in allergy; and (iii) models that study the maintenance of Tregs (homeostasis).

MODELS OF THE MECHANISMS OF TREG ACTION

An early model explicitly considering Tregs was developed by León and collaborators (92). They considered cross-regulation, where simultaneous conjugation of a Treg and an effector cell on the same APC can suppress effector function (92, 93). In this model, regulation could be due to competition for conjugation sites on the APC, or through inhibitory signals delivered to effector cells on the same APC, or by inducing conversion of effector cells to a regulatory phenotype. This model is developed and analyzed in detail in several subsequent publications (93–95), and it is reviewed in Carneiro et al. (96). Recently the model was expanded to study the dual effect of IL2 in promoting immunity and tolerance (97, 98). Some authors considered in more detail the dynamics of antigen and APCs, and compared mechanisms where regulatory T cells suppress APCs function or maturation with models where Tregs act directly on effector T cells (99). Other models along these lines included the processes of APC maturation and the differentiation of T cells into regulatory or effector phenotypes (100), following a previous proposal for this interaction (101). Interestingly, in these models, survival or proliferation of Tregs is dependent on feedback from effector T-cells, which is in part responsible for the bi-stability observed that is interpreted as states of tolerance or immunity.

Another mechanism of peripheral tolerance modeled by several authors involves anergy of effector cells (102, 103). This anergy can be achieved by tuning the threshold for activation, for example through repeated encounter with antigen or APC (102–106), or through modulation by Tregs. Carneiro et al. compared this mechanism with their previous model of cross-regulation discussed above (102). Another model that also explores thresholds for activation, but based on effector T-cell population response was studied by Burroughs et al. (107, 108). In this model the relative levels of Tregs and effector T cells depend on the respective strength of stimulation by antigen, which can be modulated by IL-2 – this model is reviewed in (109).

Typically these models consider a limited number of cell populations (3–6) and analyze one mechanism at a time. However, Kim et al. proposed a detailed model including dozens of cell populations, with a spatial component (tissue and lymph node), and considered multiple mechanisms of Treg action simultaneously (110). At the other end of the spectrum, Abreu et al. proposed a model where regulation of the immune system was simply based on cross-recognition of multiple antigens by the same cell, whether it is an effector, a regulatory, or an APC (111).

MODELS OF THE EFFECT OF TREGS ON THE IMMUNE RESPONSE

The studies discussed so far are mainly concerned with the mechanisms defining the interactions of Tregs and effector cells, often looking for steady states where one or the other population dominates, interpreted as tolerance or autoimmune states. Other models analyze the effects of the existence of Tregs on different processes.

Many of these models explore the system level effects of Treg failure and the potential development of autoimmunity. One of the first models to study this was by León et al. (112), where they analyzed the relationship between infections and autoimmunity in general. More recent studies analyzed specific autoimmune conditions, such as multiple sclerosis (113, 114) and inflammatory bowel disease (115). Grosse et al. analyzed the balance of Th1 vs. Th2 type responses and their control by Tregs in the context of allergies, with the objective of analyzing immunotherapy protocols (116). Another study looking at immunotherapy protocols, in this case modulation of IL2 therapy, used a mathematical model of helper, regulatory, and memory $CD4^+$ T cells (98). These studies are mostly theoretical. However, one report described an interesting experimental study of mice injected with tolerogenic or control peptides and followed for 16 days, with serial measurements of different T-cell subsets. These data were then analyzed with a mathematical model (117).

Some studies have analyzed the interplay between infections and regulation of the immune response. One of these modeled the immune decision between attacking or not a given antigen, based on the network of interactions between Tregs, Th17 cells, and growing levels of antigen (as in the case of a pathogen) (118). And a model analyzing the regulation of the immune response in early HIV infection, through the expansion of specific Tregs, was recently introduced (119). Finally, León et al. (120, 121) considered an expansion of their mechanistic model of Treg – T effector interactions to study the immune response against tumors, and their control or expansion.

MODELS OF TREG HOMEOSTASIS

An important question that has also been addressed by modeling studies is the maintenance of a healthy number of Tregs. Several studies included the possibility of the Treg population being maintained in the periphery in part by feedback from the effector cells (92, 99, 100). One such model (94) studied the effects of thymic output and positive/negative selection on proper balance between tolerance and immunity in the periphery, and concluded that repertoire selection plays an important role in maintaining

Table 1 | Summary of mechanistic models of Treg action.

Cell populations considered	Mechanisms of regulation of immune response	Some properties of the model	Reference
APC, Treg, Teff, and Treg, Teff conjugates on APC No explicit APC dynamics	Competition for activation on APC Tregs inhibit Teff on same conjugate Treg maintenance is dependent on Teff	Treg inhibit growth of Teff Treg induce Teff to become Treg	(92–94, 96)
As above plus IL2	Competition for IL2 Tregs condition APC	Non-local interactions Model used to study IL2-based therapies	(97, 98)
APC and Ag dynamics APC maturation T cells are activated into Treg or Teff by APC stimulation	Tregs directly suppress Teff (specifically and bystander) Tregs suppress APC maturation	Bystander effects are important Direct suppression was more effective	(99)
Antigen Immature APC, resting APC, activated APC Precursor T cells (Tp), Teff, Treg	Tp become Treg by interaction with resting APC Tp become Teff by interaction with activated APC Teff activates APC Treg induces activated APC to rest	Strength of antigen stimulus is crucial in defining whether system is in tolerant or non-tolerant state	(100)
Stochastic model of TCR triggering for T cells (both thymus and periphery)	Different thresholds for activation vs. anergy, with or without co-stimulation	Self-reactive cells in periphery are controlled by a mechanism of reversible anergy	(103)
T cells with tunable activation thresholds	Model for integration of signals in successive encounters with APC	Exhibits self-tolerance "More cells should lead to less anergy," which is not seen in adoptive transfer experiments	(102)
Inactive and active Treg and Teff IL2 for Teff proliferation, also helps Treg proliferate Cytokine (e.g., IL7) for Treg homeostasis	Tregs consume IL2 Treg inhibit Teff (from active to inactive) proportionally to Treg numbers	Strength of antigen stimulation (for Treg and Teff) defines relative levels of those two populations	(107–109)
APC with different antigens Teff of multiple specificities Tregs of multiple specificities	Cells interact with extensive cross-reactivity, but different avidities	Effector functions are the outcome of individual cellular decisions (based on cross-reactivity) A threshold of conjugation time can be identified that permits self/non-self discrimination	(111)

Comparison of some of the models for Treg action discussed in the text. The model by Kim et al. (110) is too complex to fit in this summary table.

that balance. Baltcheva et al. (122) developed a more detailed model to analyze the life-long dynamics of precursor and mature $CD25^+$ T cells (Tregs) in humans, including thymic production, density-dependent homeostasis, and effector T-cell conversion.

CONCLUSION

The field of regulatory T cells, although relatively recent, has had an explosion of knowledge driven by detailed experimental work (20, 65, 72, 123, 124). Indeed there are many more studies than we could possibly review or even allude to in this mini-review. However, the mechanistic details of this important function of the immune system are not completely elucidated (72). Many authors have developed mathematical models of the interactions between Tregs and effector cells to try to add to our understanding of these mechanisms. Still, there is a lack of true collaborations between experimental scientists and modelers in this field. Clearly, more progress would be possible if such integrated teams worked together, as has been the case in other areas of medicine, e.g., modeling of viral infections (125). As the field matures and competing models are winnowed out, it is likely that we will be able to quantify how tolerance-inducing strategies, such as CD4-blockade, affect T-cell dynamics, and what mechanisms explain the observed behavior of T-cell based tolerance.

ACKNOWLEDGMENTS

The authors thank Joana Duarte for assistance in preparing **Figure 1**. Luis Graca is funded by FCT/PTDC/SAU-TOX/114424/2009 and FCT/PTDC/SAU-IMU/120225/2010. Ruy M. Ribeiro received funding from the European Union 7th Framework Programme under grant n° PCOFUND-GA-2009-246542 and from Fundação para a Ciência e Tecnologia, Portugal.

REFERENCES

1. Medawar PB. The behaviour and fate of skin autografts and skin homografts in rabbits: a report to the War Wounds Committee of the Medical Research Council. *J Anat* (1944) **78**(Pt 5):176–99.Epub 1944/10/01
2. Kappler JW, Roehm N, Marrack P. T cell tolerance by clonal elimination in the thymus. *Cell* (1987) **49**(2):273–80. doi:10.1016/0092-8674(87)90568-X Epub 1987/04/24

3. MacDonald HR, Schneider R, Lees RK, Howe RC, Acha-Orbea H, Festenstein H, et al. T-cell receptor V beta use predicts reactivity and tolerance to Mlsa-encoded antigens. *Nature* (1988) **332**(6159):40–5. doi:10.1038/332040a0 Epub 1988/03/03
4. von Boehmer H, Teh HS, Kisielow P. The thymus selects the useful, neglects the useless and destroys the harmful. *Immunol Today* (1989) **10**(2):57–61. doi:10.1016/0167-5699(89)90307-1 Epub 1989/02/01
5. Laufer TM, DeKoning J, Markowitz JS, Lo D, Glimcher LH. Unopposed positive selection and autoreactivity in mice expressing class II MHC only on thymic cortex. *Nature* (1996) **383**(6595):81–5. doi:10.1038/383081a0 Epub 1996/09/05
6. Laufer TM, Fan L, Glimcher LH. Self-reactive T cells selected on thymic cortical epithelium are polyclonal and are pathogenic in vivo. *J Immunol* (1999) **162**(9):5078–84.Epub 1999/05/05
7. Ramsdell F, Lantz T, Fowlkes BJ. A nondeletional mechanism of thymic self tolerance. *Science* (1989) **246**(4933):1038–41. doi:10.1126/science.2511629 Epub 1989/11/24
8. Schonrich G, Momburg F, Hammerling GJ, Arnold B. Anergy induced by thymic medullary epithelium. *Eur J Immunol* (1992) **22**(7):1687–91. doi:10.1002/eji.1830220704 Epub 1992/07/01
9. Schwartz RH. Models of T cell anergy: is there a common molecular mechanism? *J Exp Med* (1996) **184**(1):1–8. doi:10.1084/jem.184.1.1 Epub 1996/07/01
10. Lamb JR, Skidmore BJ, Green N, Chiller JM, Feldmann M. Induction of tolerance in influenza virus-immune T lymphocyte clones with synthetic peptides of influenza hemagglutinin. *J Exp Med* (1983) **157**(5):1434–47. doi:10.1084/jem.157.5.1434 Epub 1983/05/01
11. Schwartz RH. A cell culture model for T lymphocyte clonal anergy. *Science* (1990) **248**(4961):1349–56. doi:10.1126/science.2113314 Epub 1990/06/15
12. Lamb JR, Feldmann M. Essential requirement for major histocompatibility complex recognition in T-cell tolerance induction. *Nature* (1984) **308**(5954):72–4. doi:10.1038/308072a0 Epub 1984/03/01
13. Jenkins MK, Schwartz RH. Antigen presentation by chemically modified splenocytes induces antigen-specific T cell unresponsiveness in vitro and in vivo. *J Exp Med* (1987) **165**(2):302–19. doi:10.1084/jem.165.2.302 Epub 1987/02/01
14. Quill H, Schwartz RH. Stimulation of normal inducer T cell clones with antigen presented by purified Ia molecules in planar lipid membranes: specific induction of a long-lived state of proliferative nonresponsiveness. *J Immunol* (1987) **138**(11):3704–12.Epub 1987/06/01
15. Lombardi G, Sidhu S, Batchelor R, Lechler R. Anergic T cells as suppressor cells in vitro. *Science* (1994) **264**(5165):1587–9. doi:10.1126/science.8202711 Epub 1994/06/10
16. Chai JG, Bartok I, Chandler P, Vendetti S, Antoniou A, Dyson J, et al. Anergic T cells act as suppressor cells in vitro and in vivo. *Eur J Immunol* (1999) **29**(2):686–92. doi:10.1002/(SICI)1521-4141(199902)29:02<686::AID-IMMU686>3.0.CO;2-N Epub 1999/03/04
17. Vendetti S, Chai JG, Dyson J, Simpson E, Lombardi G, Lechler R. Anergic T cells inhibit the antigen-presenting function of dendritic cells. *J Immunol* (2000) **165**(3):1175–81.Epub 2000/07/21
18. Waldmann H, Qin S, Cobbold S. Monoclonal antibodies as agents to reinduce tolerance in autoimmunity. *J Autoimmun* (1992) **5**(Suppl A):93–102. doi:10.1016/0896-8411(92)90024-K Epub 1992/04/01
19. Sakaguchi S, Sakaguchi N, Asano M, Itoh M, Toda M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J Immunol* (1995) **155**(3):1151–64.Epub 1995/08/01
20. Ohkura N, Kitagawa Y, Sakaguchi S. Development and maintenance of regulatory T cells. *Immunity* (2013) **38**(3):414–23. doi:10.1016/j.immuni.2013.03.002
21. Yagi H, Nomura T, Nakamura K, Yamazaki S, Kitawaki T, Hori S, et al. Crucial role of FOXP3 in the development and function of human CD25+CD4+ regulatory T cells. *Int Immunol* (2004) **16**(11):1643–56. doi:10.1093/intimm/dxh165
22. Fontenot JD, Gavin MA, Rudensky AY. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. *Nat Immunol* (2003) **4**(4):330–6. doi:10.1038/ni904
23. Hori S, Nomura T, Sakaguchi S. Control of regulatory T cell development by the transcription factor Foxp3. *Science* (2003) **299**(5609):1057–61. doi:10.1126/science.1079490 Epub 2003/01/11
24. Khattri R, Cox T, Yasayko SA, Ramsdell F. An essential role for Scurfin in CD4+CD25+ T regulatory cells. *Nat Immunol* (2003) **4**(4):337–42. doi:10.1038/ni909
25. Sakaguchi S. Naturally arising CD4+ regulatory t cells for immunologic self-tolerance and negative control of immune responses. *Annu Rev Immunol* (2004) **22**:531–62. doi:10.1146/annurev.immunol.21.120601.141122 Epub 2004/03/23
26. Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, Whitesell L, et al. The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nat Genet* (2001) **27**(1):20–1. doi:10.1038/83713 Epub 2001/01/04
27. Wildin RS, Ramsdell F, Peake J, Faravelli F, Casanova JL, Buist N, et al. X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. *Nat Genet* (2001) **27**(1):18–20. doi:10.1038/83707 Epub 2001/01/04
28. Brunkow ME, Jeffery EW, Hjerrild KA, Paeper B, Clark LB, Yasayko SA, et al. Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. *Nat Genet* (2001) **27**(1):68–73. doi:10.1038/83784
29. Apostolou I, Sarukhan A, Klein L, von Boehmer H. Origin of regulatory T cells with known specificity for antigen. *Nat Immunol* (2002) **3**(8):756–63. doi:10.1038/ni816
30. Jordan MS, Boesteanu A, Reed AJ, Petrone AL, Holenbeck AE, Lerman MA, et al. Thymic selection of CD4+CD25+ regulatory T cells induced by an agonist self-peptide. *Nat Immunol* (2001) **2**(4):301–6. doi:10.1038/86302 Epub 2001/03/29
31. Tai X, Cowan M, Feigenbaum L, Singer A. CD28 costimulation of developing thymocytes induces Foxp3 expression and regulatory T cell differentiation independently of interleukin 2. *Nat Immunol* (2005) **6**(2):152–62. doi:10.1038/ni1160 Epub 2005/01/11
32. Fontenot JD, Rasmussen JP, Williams LM, Dooley JL, Farr AG, Rudensky AY. Regulatory T cell lineage specification by the forkhead transcription factor foxp3. *Immunity* (2005) **22**(3):329–41. doi:10.1016/j.immuni.2005.01.016
33. Fontenot JD, Rasmussen JP, Gavin MA, Rudensky AY. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nat Immunol* (2005) **6**(11):1142–51. doi:10.1038/ni1263
34. Marie JC, Letterio JJ, Gavin M, Rudensky AY. TGF-beta1 maintains suppressor function and Foxp3 expression in CD4+CD25+ regulatory T cells. *J Exp Med* (2005) **201**(7):1061–7. doi:10.1084/jem.20042276
35. Gavin MA, Clarke SR, Negrou E, Gallegos A, Rudensky A. Homeostasis and anergy of CD4(+)CD25(+) suppressor T cells in vivo. *Nat Immunol* (2002) **3**(1):33–41. doi:10.1038/ni743
36. Knoechel B, Lohr J, Kahn E, Bluestone JA, Abbas AK. Sequential development of interleukin 2-dependent effector and regulatory T cells in response to endogenous systemic antigen. *J Exp Med* (2005) **202**(10):1375–86. doi:10.1084/jem.20050855
37. Lohr J, Knoechel B, Jiang S, Sharpe AH, Abbas AK. The inhibitory function of B7 costimulators in T cell responses to foreign and self-antigens. *Nat Immunol* (2003) **4**(7):664–9. doi:10.1038/ni939
38. Malek TR. The main function of IL-2 is to promote the development of T regulatory cells. *J Leukoc Biol* (2003) **74**(6):961–5. doi:10.1189/jlb.0603272
39. Josefowicz SZ, Lu LF, Rudensky AY. Regulatory T cells: mechanisms of differentiation and function. *Annu Rev Immunol* (2012) **30**:531–64. doi:10.1146/annurev.immunol.25.022106.141623 Epub 2012/01/10
40. Salomon B, Lenschow DJ, Rhee L, Ashourian N, Singh B, Sharpe A, et al. B7/CD28 costimulation is essential for the homeostasis of the CD4+CD25+ immunoregulatory T cells that control autoimmune diabetes. *Immunity* (2000) **12**(4):431–40. doi:10.1016/S1074-7613(00)80195-8 Epub 2000/05/05
41. Graca L, Le Moine A, Lin CY, Fairchild PJ, Cobbold SP, Waldmann H. Donor-specific transplantation tolerance: the paradoxical behavior of CD4+CD25+ T cells. *Proc Natl Acad Sci U S A* (2004) **101**(27):10122–6. doi:10.1073/pnas.0400084101 Epub 2004/06/26
42. Davies JD, Leong LY, Mellor A, Cobbold SP, Waldmann H. T cell suppression in transplantation tolerance through linked recognition. *J Immunol* (1996) **156**(10):3602–7.Epub 1996/05/15
43. Qin S, Cobbold SP, Pope H, Elliott J, Kioussis D, Davies J, et al. "Infectious" transplantation tolerance. *Science* (1993) **259**(5097):974–7.Epub 1993/02/12

44. Graca L, Honey K, Adams E, Cobbold SP, Waldmann H. Cutting edge: anti-CD154 therapeutic antibodies induce infectious transplantation tolerance. *J Immunol* (2000) **165**(9):4783–6.Epub 2000/10/25
45. Kendal AR, Chen Y, Regateiro FS, Ma J, Adams E, Cobbold SP, et al. Sustained suppression by Foxp3+ regulatory T cells is vital for infectious transplantation tolerance. *J Exp Med* (2011) **208**(10):2043–53. doi:10.1084/jem.20110767 Epub 2011/08/31
46. Apostolou I, von Boehmer H. In vivo instruction of suppressor commitment in naive T cells. *J Exp Med* (2004) **199**(10):1401–8. doi:10.1084/jem.20040249 Epub 2004/05/19
47. Kretschmer K, Apostolou I, Hawiger D, Khazaie K, Nussenzweig MC, von Boehmer H. Inducing and expanding regulatory T cell populations by foreign antigen. *Nat Immunol* (2005) **6**(12):1219–27. doi:10.1038/ni1265 Epub 2005/10/26
48. Cobbold SP, Castejon R, Adams E, Zelenika D, Graca L, Humm S, et al. Induction of foxP3+ regulatory T cells in the periphery of T cell receptor transgenic mice tolerized to transplants. *J Immunol* (2004) **172**(10):6003–10.Epub 2004/05/07
49. Curotto de Lafaille MA, Lino AC, Kutchukhidze N, Lafaille JJ. CD25− T cells generate CD25+Foxp3+ regulatory T cells by peripheral expansion. *J Immunol* (2004) **173**(12):7259–68.
50. Furtado GC, Curotto de Lafaille MA, Kutchukhidze N, Lafaille JJ. Interleukin 2 signaling is required for CD4(+) regulatory T cell function. *J Exp Med* (2002) **196**(6):851–7. doi:10.1084/jem.20020190
51. Cobbold SP, Graca L, Lin CY, Adams E, Waldmann H. Regulatory T cells in the induction and maintenance of peripheral transplantation tolerance. *Transpl Int* (2003) **16**(2):66–75. doi:10.1111/j.1432-2277.2003.tb00266.x Epub 2003/02/22
52. Liang S, Alard P, Zhao Y, Parnell S, Clark SL, Kosiewicz MM. Conversion of CD4+ CD25− cells into CD4+ CD25+ regulatory T cells in vivo requires B7 costimulation, but not the thymus. *J Exp Med* (2005) **201**(1):127–37. doi:10.1084/jem.20041201
53. Chen W, Jin W, Hardegen N, Lei KJ, Li L, Marinos N, et al. Conversion of peripheral CD4+CD25− naive T cells to CD4+CD25+ regulatory T cells by TGF-beta induction of transcription factor Foxp3. *J Exp Med* (2003) **198**(12):1875–86. doi:10.1084/jem.20030152
54. Selvaraj RK, Geiger TL. A kinetic and dynamic analysis of Foxp3 induced in T cells by TGF-beta. *J Immunol* (2007) **178**(12):7667–77.
55. Zheng SG, Wang JH, Gray JD, Soucier H, Horwitz DA. Natural and induced CD4+CD25+ cells educate CD4+CD25− cells to develop suppressive activity: the role of IL-2, TGF-beta, and IL-10. *J Immunol* (2004) **172**(9):5213–21.
56. Graca L, Chen TC, Le Moine A, Cobbold SP, Howie D, Waldmann H. Dominant tolerance: activation thresholds for peripheral generation of regulatory T cells. *Trends Immunol* (2005) **26**(3):130–5. doi:10.1016/j.it.2004.12.007 Epub 2005/03/05
57. Oliveira VG, Caridade M, Paiva RS, Demengeot J, Graca L. Sub-optimal CD4+ T-cell activation triggers autonomous TGF-beta-dependent conversion to Foxp3+ regulatory T cells. *Eur J Immunol* (2011) **41**(5):1249–55. doi:10.1002/eji.201040896 Epub 2011/04/07
58. Mucida D, Kutchukhidze N, Erazo A, Russo M, Lafaille JJ, Curotto de Lafaille MA. Oral tolerance in the absence of naturally occurring Tregs. *J Clin Invest* (2005) **115**(7):1923–33. doi:10.1172/JCI24487 Epub 2005/06/07
59. Lin CY, Graca L, Cobbold SP, Waldmann H. Dominant transplantation tolerance impairs CD8+ T cell function but not expansion. *Nat Immunol* (2002) **3**(12):1208–13. doi:10.1038/ni853 Epub 2002/11/05
60. Chen TC, Waldmann H, Fairchild PJ. Induction of dominant transplantation tolerance by an altered peptide ligand of the male antigen Dby. *J Clin Invest* (2004) **113**(12):1754–62. doi:10.1172/JCI20569 Epub 2004/06/17
61. Griffin MD, Lutz W, Phan VA, Bachman LA, McKean DJ, Kumar R. Dendritic cell modulation by 1alpha,25 dihydroxyvitamin D3 and its analogs: a vitamin D receptor-dependent pathway that promotes a persistent state of immaturity in vitro and in vivo. *Proc Natl Acad Sci U S A* (2001) **98**(12):6800–5. doi:10.1073/pnas.121172198 Epub 2001/05/24
62. Manicassamy S, Pulendran B. Dendritic cell control of tolerogenic responses. *Immunol Rev* (2011) **241**(1):206–27. doi:10.1111/j.1600-065X.2011.01015.x Epub 2011/04/15
63. Mucida D, Park Y, Kim G, Turovskaya O, Scott I, Kronenberg M, et al. Reciprocal TH17 and regulatory T cell differentiation mediated by retinoic acid. *Science* (2007) **317**(5835):256–60. doi:10.1126/science.1145697 Epub 2007/06/16
64. Steinbrink K, Wolfl M, Jonuleit H, Knop J, Enk AH. Induction of tolerance by IL-10-treated dendritic cells. *J Immunol* (1997) **159**(10):4772–80.Epub 1997/11/20
65. Waldmann H, Chen TC, Graca L, Adams E, Daley S, Cobbold S, et al. Regulatory T cells in transplantation. *Semin Immunol* (2006) **18**(2):111–9. doi:10.1016/j.smim.2006.01.010 Epub 2006/02/16
66. Yang J, Bernier SM, Ichim TE, Li M, Xia X, Zhou D, et al. LF15-0195 generates tolerogenic dendritic cells by suppression of NF-kappaB signaling through inhibition of IKK activity. *J Leukoc Biol* (2003) **74**(3):438–47. doi:10.1189/jlb.1102582 Epub 2003/09/02
67. Battaglia M, Stabilini A, Migliavacca B, Horejs-Hoeck J, Kaupper T, Roncarolo MG. Rapamycin promotes expansion of functional CD4+CD25+FOXP3+ regulatory T cells of both healthy subjects and type 1 diabetic patients. *J Immunol* (2006) **177**(12):8338–47.
68. Cobbold SP, Adams E, Farquhar CA, Nolan KF, Howie D, Lui KO, et al. Infectious tolerance via the consumption of essential amino acids and mTOR signaling. *Proc Natl Acad Sci U S A* (2009) **106**(29):12055–60. doi:10.1073/pnas.0903919106
69. Zhang R, Huynh A, Whitcher G, Chang J, Maltzman JS, Turka LA. An obligate cell-intrinsic function for CD28 in Tregs. *J Clin Invest* (2013) **123**(2):580–93. doi:10.1172/JCI65013
70. Miyara M, Sakaguchi S. Natural regulatory T cells: mechanisms of suppression. *Trends Mol Med* (2007) **13**(3):108–16. doi:10.1016/j.molmed.2007.01.003 Epub 2007/01/30
71. Toda A, Piccirillo CA. Development and function of naturally occurring CD4+CD25+ regulatory T cells. *J Leukoc Biol* (2006) **80**(3):458–70. doi:10.1189/jlb.0206095 Epub 2006/07/01
72. Shevach EM. Mechanisms of foxp3+ T regulatory cell-mediated suppression. *Immunity* (2009) **30**(5):636–45. doi:10.1016/j.immuni.2009.04.010
73. Wing K, Fehervari Z, Sakaguchi S. Emerging possibilities in the development and function of regulatory T cells. *Int Immunol* (2006) **18**(7):991–1000. doi:10.1093/intimm/dxl044 Epub 2006/05/25
74. Grossman WJ, Verbsky JW, Barchet W, Colonna M, Atkinson JP, Ley TJ. Human T regulatory cells can use the perforin pathway to cause autologous target cell death. *Immunity* (2004) **21**(4):589–601. doi:10.1016/j.immuni.2004.09.002 Epub 2004/10/16
75. Gondek DC, Lu LF, Quezada SA, Sakaguchi S, Noelle RJ. Cutting edge: contact-mediated suppression by CD4+CD25+ regulatory cells involves a granzyme B-dependent, perforin-independent mechanism. *J Immunol* (2005) **174**(4):1783–6.Epub 2005/02/09.
76. Sakaguchi S, Wing K, Onishi Y, Prieto-Martin P, Yamaguchi T. Regulatory T cells: how do they suppress immune responses? *Int Immunol* (2009) **21**(10):1105–11. doi:10.1093/intimm/dxp095 Epub 2009/09/10
77. Vignali DA, Collison LW, Workman CJ. How regulatory T cells work. *Nat Rev Immunol* (2008) **8**(7):523–32. doi:10.1038/nri2343
78. Kryczek I, Wei S, Zou L, Zhu G, Mottram P, Xu H, et al. Cutting edge: induction of B7-H4 on APCs through IL-10: novel suppressive mode for regulatory T cells. *J Immunol* (2006) **177**(1):40–4.Epub 2006/06/21
79. Scheffold A, Huhn J, Hofer T. Regulation of CD4+CD25+ regulatory T cell activity: it takes (IL-)two to tango. *Eur J Immunol* (2005) **35**(5):1336–41. doi:10.1002/eji.200425887 Epub 2005/04/14
80. Scheffold A, Murphy KM, Hofer T. Competition for cytokines: T(reg) cells take all. *Nat Immunol* (2007) **8**(12):1285–7. doi:10.1038/ni1207-1285 Epub 2007/11/21
81. Maynard CL, Harrington LE, Janowski KM, Oliver JR, Zindl CL, Rudensky AY, et al. Regulatory T cells expressing interleukin 10 develop from Foxp3+ and Foxp3− precursor cells in the absence of interleukin 10. *Nat Immunol* (2007) **8**(9):931–41. doi:10.1038/ni1504
82. Rubtsov YP, Rasmussen JP, Chi EY, Fontenot J, Castelli L, Ye X, et al. Regulatory T cell-derived interleukin-10 limits inflammation at environmental interfaces. *Immunity* (2008) **28**(4):546–58. doi:10.1016/j.immuni.2008.02.017
83. Cao X, Cai SF, Fehniger TA, Song J, Collins LI, Piwnica-Worms DR, et al. Granzyme B and perforin are important for regulatory T cell-mediated suppression of tumor clearance. *Immunity* (2007) **27**(4):635–46. doi:10.1016/j.immuni.2007.08.014
84. Faro J, Velasco S, Gonzalez-Fernandez A, Bandeira A. The impact of thymic antigen diversity on the size of the selected T cell repertoire. *J Immunol* (2004) **172**(4):2247–55.

85. Detours V, Mehr R, Perelson AS. Deriving quantitative constraints on T cell selection from data on the mature T cell repertoire. *J Immunol* (2000) **164**(1):121–8.
86. Detours V, Perelson AS. Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proc Natl Acad Sci U S A* (1999) **96**(9):5153–8. doi:10.1073/pnas.96.9.5153
87. Mehr R, Globerson A, Perelson AS. Modeling positive and negative selection and differentiation processes in the thymus. *J Theor Biol* (1995) **175**(1):103–26. doi:10.1006/jtbi.1995.0124
88. Currie J, Castro M, Lythe G, Palmer E, Molina-Paris C. A stochastic T cell response criterion. *J R Soc Interface* (2012) **9**(76):2856–70. doi:10.1098/rsif.2012.0205
89. Castiglione F, Santoni D, Rapin N. CTLs' repertoire shaping in the thymus: a Monte Carlo simulation. *Autoimmunity* (2011) **44**(4):261–70. doi:10.3109/08916934.2011.523272
90. Muller V, Bonhoeffer S. Quantitative constraints on the scope of negative selection. *Trends Immunol* (2003) **24**(3):132–5. doi:10.1016/S1471-4906(03)00028-0
91. Sawicka ML, Reynolds J, Abourashchi N, Lythe G, Molina-Paris C, et al. Immunology and mathematics: a joint effort to estimate (murine) thymic selection rates. *Front Immunol* (in press).
92. Leon K, Perez R, Lage A, Carneiro J. Modelling T-cell-mediated suppression dependent on interactions in multicellular conjugates. *J Theor Biol* (2000) **207**(2):231–54. doi:10.1006/jtbi.2000.2169 Epub 2000/10/18
93. Leon K, Perez R, Lage A, Carneiro J. Three-cell interactions in T cell-mediated suppression? A mathematical analysis of its quantitative implications. *J Immunol* (2001) **166**(9):5356–65.Epub 2001/04/21
94. Leon K, Lage A, Carneiro J. Tolerance and immunity in a mathematical model of T-cell mediated suppression. *J Theor Biol* (2003) **225**(1):107–26. doi:10.1016/S0022-5193(03)00226-1 Epub 2003/10/16
95. Leon K, Faro J, Carneiro J. A general mathematical framework to model generation structure in a population of asynchronously dividing cells. *J Theor Biol* (2004) **229**(4):455–76. doi:10.1016/j.jtbi.2004.04.011 Epub 2004/07/13
96. Carneiro J, Leon K, Caramalho I, van den Dool C, Gardner R, Oliveira V, et al. When three is not a crowd: a crossregulation model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunol Rev* (2007) **216**:48–68. doi:10.1111/j.1600-065X.2007.00487.x. Epub 2007/03/21
97. Garcia-Martinez K, Leon K. Modeling the role of IL-2 in the interplay between CD4+ helper and regulatory T cells: assessing general dynamical properties. *J Theor Biol* (2010) **262**(4):720–32. doi:10.1016/j.jtbi.2009.10.025 Epub 2009/11/03
98. Garcia-Martinez K, Leon K. Modeling the role of IL2 in the interplay between CD4+ helper and regulatory T cells: studying the impact of IL2 modulation therapies. *Int Immunol* (2012) **24**(7):427–46. doi:10.1093/intimm/dxr120
99. Alexander HK, Wahl LM. Self-tolerance and autoimmunity in a regulatory T cell model. *Bull Math Biol* (2011) **73**(1):33–71. doi:10.1007/s11538-010-9519-2 Epub 2010/03/03
100. Fouchet D, Regoes R. A population dynamics analysis of the interaction between adaptive regulatory T cells and antigen presenting cells. *PLoS One* (2008) **3**(5):e2306. doi:10.1371/journal.pone.0002306. Epub 2008/05/30
101. Powrie F, Maloy KJ. Immunology. regulating the regulators. *Science* (2003) **299**(5609):1030–1. doi:10.1126/science.1082031
102. Carneiro J, Paixao T, Milutinovic D, Sousa J, Leon K, Gardner R, et al. Immunological self-tolerance: Lessons from mathematical modeling. *J Comput Appl Math* (2005) **184**(1):77–100. doi:10.1016/j.cam.2004.10.025
103. Chan C, Stark J, George AJT. The impact of multiple T cell-APC encounters and the role of anergy. *J Comput Appl Math* (2005) **184**(1):101–20. doi:10.1016/j.cam.2004.07.036
104. Anderton SM, Wraith DC. Selection and fine-tuning of the autoimmune T-cell repertoire. *Nat Rev Immunol* (2002) **2**(7):487–98. doi:10.1038/nri842
105. Grossman Z, Paul WE. Adaptive cellular interactions in the immune system: the tunable activation threshold and the significance of subthreshold responses. *Proc Natl Acad Sci U S A* (1992) **89**(21):10365–9. doi:10.1073/pnas.89.21.10365
106. van den Berg HA, Rand DA. Quantitative theories of T-cell responsiveness. *Immunol Rev* (2007) **216**:81–92. doi:10.1111/j.1600-065X.2006.00491.x
107. Burroughs NJ, Miguel Paz Mendes de Oliveira B, Adrego Pinto A. Regulatory T cell adjustment of quorum growth thresholds and the control of local immune responses. *J Theor Biol* (2006) **241**(1):134–41. doi:10.1016/j.jtbi.2005.11.010 Epub 2006/01/13
108. Burroughs NJ, Oliveira BM, Pinto AA, Sequeira HJT. Sensibility of the quorum growth thresholds controlling local immune responses. *Math Comput Model* (2008) **47**(7–8):714–25. doi:10.1016/j.mcm.2007.06.007
109. Pinto AA, Burroughs NJ, Ferreira M, Oliveira BM. Dynamics of immunological models. *Acta Biotheor* (2010) **58**(4):391–404. doi:10.1007/s10441-010-9117-6
110. Kim PS, Lee PP, Levy D. Modeling regulation mechanisms in the immune system. *J Theor Biol* (2007) **246**(1):33–69. doi:10.1016/j.jtbi.2006.12.012 Epub 2007/02/03
111. de Abreu FV, Mostardinha P. Maximal frustration as an immunological principle. *J R Soc Interface* (2009) **6**(32):321–34. doi:10.1098/rsif.2008.0280
112. Leon K, Faro J, Lage A, Carneiro J. Inverse correlation between the incidences of autoimmune disease and infection predicted by a model of T cell mediated tolerance. *J Autoimmun* (2004) **22**(1):31–42. doi:10.1016/j.jaut.2003.10.002 Epub 2004/01/08
113. Martinez-Pasamar S, Abad E, Moreno B, Velez de Mendizabal N, Martinez-Forero I, Garcia-Ojalvo J, et al. Dynamic cross-regulation of antigen-specific effector and regulatory T cell subpopulations and microglia in brain autoimmunity. *BMC Syst Biol* (2013) **7**:34. doi:10.1186/1752-0509-7-34
114. Velez de Mendizabal N, Carneiro J, Sole RV, Goni J, Bragard J, Martinez-Forero I, et al. Modeling the effector – regulatory T cell cross-regulation reveals the intrinsic character of relapses in Multiple Sclerosis. *BMC Syst Biol* (2011) **5**:114. doi:10.1186/1752-0509-5-114
115. Lo WC, Arsenescu RI, Friedman A. Mathematical Model of the Roles of T Cells in Inflammatory Bowel Disease. *Bull Math Biol* (2013) **75**(9):1417–33. doi:10.1007/s11538-013-9853-2
116. Gross F, Metzner G, Behn U. Mathematical modeling of allergy and specific immunotherapy: Th1-Th2-Treg interactions. *J Theor Biol* (2011) **269**(1):70–8. doi:10.1016/j.jtbi.2010.10.013
117. Arazi A, Sharabi A, Zinger H, Mozes E, Neumann AU. In vivo dynamical interactions between CD4 Tregs, CD8 Tregs and CD4+ CD25− cells in mice. *PLoS One* (2009) **4**(12):e8447. doi:10.1371/journal.pone.0008447
118. Bewick S, Yang R, Zhang M. The danger is growing! A new paradigm for immune system activation and peripheral tolerance. *PLoS One* (2009) **4**(12):e8112. doi:10.1371/journal.pone.0008112
119. Simonov M, Rawlings RA, Comment N, Reed SE, Shi X, Nelson PW. Modeling adaptive regulatory T-cell dynamics during early HIV infection. *PLoS One* (2012) **7**(4):e33924. doi:10.1371/journal.pone.0033924
120. Leon K, Garcia K, Carneiro J, Lage A. How regulatory CD25+CD4+ T cells impinge on tumor immunobiology: the differential response of tumors to therapies. *J Immunol* (2007) **179**(9):5659–68.
121. Leon K, Garcia K, Carneiro J, Lage A. How regulatory CD25(+)CD4(+) T cells impinge on tumor immunobiology? On the existence of two alternative dynamical classes of tumors. *J Theor Biol* (2007) **247**(1):122–37. doi:10.1016/j.jtbi.2007.01.029
122. Baltcheva I, Codarri L, Pantaleo G, Le Boudec JY. Lifelong dynamics of human CD4+CD25+ regulatory T cells: insights from in vivo data and mathematical modeling. *J Theor Biol* (2010) **266**(2):307–22. doi:10.1016/j.jtbi.2010.06.024
123. Agua-Doce A, Graca L. Regulatory T cells and the control of the allergic response. *J Allergy* (2012) **2012**:948901. doi:10.1155/2012/948901 Epub 2012/10/12
124. Gratz IK, Rosenblum MD, Abbas AK. The life of regulatory T cells. *Ann N Y Acad Sci* (2013) **1283**:8–12. doi:10.1111/nyas.12011
125. Perelson AS. Modelling viral and immune system dynamics. *Nat Rev Immunol* (2002) **2**(1):28–36. doi:10.1038/nri700

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 August 2013; paper pending published: 23 September 2013; accepted: 03 November 2013; published online: 18 November 2013.
*Citation: Caridade M, Graca L and Ribeiro RM (2013) Mechanisms underlying CD4+ Treg immune regulation in the adult: from experiments to models. Front. Immunol. **4**:378. doi: 10.3389/fimmu.2013.00378*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Caridade, Graca and Ribeiro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 November 2013
doi: 10.3389/fimmu.2013.00380

Mathematical modeling of oncogenesis control in mature T-cell populations

Sebastian Gerdes[1], Sebastian Newrzela[2], Ingmar Glauche[1], Dorothee von Laer[3], Martin-Leo Hansmann[2] and Ingo Roeder[1]*

[1] *Institute for Medical Informatics and Biometry, Medical Faculty Carl Gustav Carus, Dresden, Germany*
[2] *Senckenberg Institute of Pathology, Goethe-University Hospital Frankfurt, Frankfurt, Germany*
[3] *Department of Hygiene, Medical University Innsbruck, Innsbruck, Austria*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Antonio A. Freitas, Institut Pasteur, France
Tomasz Zal, University of Texas MD Anderson Cancer Center, USA

****Correspondence:***
Ingo Roeder, Institute for Medical Informatics and Biometry, Medical Faculty Carl Gustav Carus, TU Dresden, Fetscherstr. 74, 01307 Dresden, Germany
e-mail: ingo.roeder@tu-dresden.de

T-cell receptor (TCR) polyclonal mature T cells are surprisingly resistant to oncogenic transformation after retroviral insertion of T-cell oncogenes. In a mouse model, it has been shown that mature T-cell lymphoma/leukemia (MTCLL) is not induced upon transplantation of mature, TCR polyclonal wild-type (WT) T cells, transduced with gammaretroviral vectors encoding potent T-cell oncogenes, into RAG1-deficient recipients. However, further studies demonstrated that quasi-monoclonal T cells treated with the same protocol readily induced MTCLL in the recipient mice. It has been hypothesized that in the TCR polyclonal situation, outgrowth of preleukemic cells and subsequent conversion to overt malignancy is suppressed through regulation of clonal abundances on a per-clone basis due to interactions between TCRs and self-peptide-MHC-complexes (spMHCs), while these mechanisms fail in the quasi-monoclonal situation. To quantitatively study this hypothesis, we applied a mathematical modeling approach. In particular, we developed a novel ordinary differential equation model of T-cell homeostasis, in which T-cell fate depends on spMHC-TCR-interaction-triggered stimulatory signals from antigen-presenting cells (APCs). Based on our mathematical modeling approach, we identified parameter configurations of our model, which consistently explain the observed phenomena. Our results suggest that the preleukemic cells are less competent than healthy competitor cells in acquiring survival stimuli from APCs, but that proliferation of these preleukemic cells is less dependent on survival stimuli from APCs. These predictions now call for experimental validation.

Keywords: T-cell homeostasis, T-cell niche, gene therapy, mature T-cell lymphoma, MTCL

1. INTRODUCTION

Mature T cells are an essential component of the adaptive immune system. They carry the so-called *T-cell receptor* (TCR) on their surface. This receptor enables them to recognize peptides that are presented to them via major histocompatibility complex (MHC) molecules on antigen-presenting cells (APCs). A vast number of different TCRs is expressed in T cells in healthy individuals, estimated to be in the order of 10^6 in mice (1) and 10^7 in humans (2). An individual T cell expresses a single TCR variant, and passes this variant on to its daughter cells. The set of all T cells expressing the same TCR is called a *T-cell clone* (or simply referred to as *clone*). The enormous TCR diversity is created during T-cell maturation in the thymus through genomic rearrangement of the TCR gene locus (3). A series of selection processes ensures that mature T cells can bind with low to moderate affinity to self-peptide-MHC complexes (spMHCs) on APCs (4–6). After maturation, T cells enter the peripheral T-cell pool.

The peripheral T-cell pool is remarkably stable in terms of cell numbers and clonal diversity throughout the lifetime of mice and humans. In order to explain this stability, concepts have emerged that are based on competition between T cells for limiting trophic resources needed for survival and proliferation (7). The limiting trophic resources can be divided in public and TCR-specific resources (8). In principle, public trophic resources are equally accessible to all T cells, and include stimulatory cytokines (e.g. interleukin 7), nutrients, costimulatory molecules, or physical space. In contrast, access to TCR-specific resources depends on the particular TCR that is expressed on a T cell. TCR-specific resources are represented mainly by stimulatory interactions with APCs due to binding of the TCR to spMHCs (9–11).

Different spMHCs may vary substantially in their suitability to mediate a stimulatory interaction for particular T-cell clones. Consequently, a *T-cell niche* concept has been proposed, in which different spMHCs represent distinct T-cell niches (12). The niches provide vital resources that different T-cell clones compete for. A particular clone may not receive resources from all niches equally well. This concept implies that the TCR diversity is stabilized by the diversity of the available spMHCs (13).

When the regulation of cellular proliferation in the T-cell system is corrupted, mature T-cell lymphoma/leukemia (MTCLL) formation may occur. However, oncogenesis is comparatively rare in mature T cells. For example, the incidence of B-cell lymphoid neoplasms is substantially higher than the incidence of T-cell/natural killer cell lymphoid neoplasms, as shown in a study from the United States ($26.13/10^5$/year vs. $1.79/10^5$/year (14)). Furthermore, several studies from the field of retroviral gene

therapy confirm the relative resistance of mature T cells to oncogenesis. Despite long follow-up times, retroviral vector-induced oncogenesis has never been observed in clinical gene therapy trials involving gene-modified mature T cells (15–17). In contrast, genotoxicity was observed in several studies involving retroviral gene transfer into hematopoietic stem and progenitor cells (HSPCs) (18, 19).

Motivated by these observations, we here focus on the analysis of oncogenesis control in mature T-cell populations.

In order to explicitly investigate the relative resistance of mature T cells to malignant transformation in a gene therapeutic context, HSPCS, and mature T cells were exposed to an identical transformation assay in a defined experimental setting (20). In this assay, HSPCs and mature T cells were isolated from wild-type mice and were each transduced independently with high copy numbers of gammaretroviral vectors encoding potent T-cell oncogenes. Subsequently, the cells were transplanted into immunoincompetent RAG1-deficient mice. HSPC-transplanted animals consistently developed MTCLL. In contrast, MTCLL has not been observed in any of the recipients that were transplanted with mature T cells. This finding corroborated the relative resistance of mature T cells to malignant transformation.

In a subsequent study, the impact of TCR diversity on T-cell resistance to malignant transformation has been further assessed (21). In this study, T-cell populations were isolated from OT1- or P14-mice, i.e. mice expressing a transgenic TCR. T-cell populations from these mouse models are quasi-monoclonal, i.e. they express predominantly one specific TCR. By applying a similar, yet refined, transformation assay as in the previous study MTCLL readily developed in the recipient RAG1-deficient mice (see **Figure 1**). Moreover, addition of untransduced TCR polyclonal T cells to quasi-monoclonal, transduced cell populations prevented malignancy development, demonstrating that TCR polyclonality plays a pivotal role in malignancy control in mature T cells.

Building on these observations, we hypothesize that in the TCR polyclonal situation, prohomeostatic signals, due to interactions between TCRs and spMHCs, suppress the outgrowth of preleukemic T cells (i.e. in this context, T cells that have been transformed by retroviral insertion of an oncogene), while these mechanisms fail in the TCR quasi-monoclonal situation, and vigorous cell expansion occurs. In this paper, we aim to quantitatively assess the implications of this hypothesis using a mathematical modeling approach. Specifically, we develop a mathematical model of a niche-dependent mature T-cell regulation [similar to previous published models, e.g. Ref. (22)], which will be applied to model the physiological situation and MTCLL formation. Using this model of T-cell homeostasis, we present an *in silico* simulation scenario that is suited to mimic the experimental procedures performed by Newrzela et al. (20, 21). In an extensive parameter screen, we evaluate if, and under which parameter constellations the experimental observations can be explained.

2. MATERIALS AND METHODS

2.1. MODEL DESCRIPTION

The two major entities in our model are *T-cell species* (also called simply *species*) and *T-cell niches* (also called simply *niches*). With the term *species*, we refer to a set of T cells that is homogeneous in terms of our model parameters. In the physiological situation, a T-cell clone can be represented as a particular species. However, in order to model the aforementioned experimental situation (20, 21), we will represent each T-cell clone by two species, namely a species representing healthy cells within a clone and a preleukemic species representing cells that potentially give rise to MTCLL.

Our model is constructed to describe the temporal dynamics of species abundances, i.e. the number of cells belonging to a particular T-cell species at any point of time. The number of species in the system is denoted by the symbol m. The species abundances at time t are represented by the vector $\mathbf{c}(t) = (c_t(t))$ with $i = 1, 2, \ldots, m$. The initial species abundances $\mathbf{c}(0)$ are defined by an m-dimensional vector denoted $\mathbf{c}^0 = \left(c_i^0\right)$.

In the model, T-cell species compete for resources needed for survival and proliferation that are supplied by T-cell niches. The number of niches in the system is denoted by the symbol n. It is assumed that the niches supply resources at constant rates, represented by the n-dimensional vector $\mathbf{p} = (p_j)$, where $j = 1, 2, \ldots, n$.

The competition for niche resources between species is defined by the $m \times n$ matrix $\mathbf{A} = [a_{ij}]$, called *niche affinity matrix*, in which a_{ij} is a measure of the capability of species i to acquire resources from niche j.

At any time point t, species receive resources from niches according to instantaneous rates. These rates, which are termed *resource acquisition rates*, are stored in an $m \times n$ matrix $\mathbf{R}(t) = [r_{ij}(t)]$. $r_{ij}(t)$ denotes the rate at which the i-th species receives resources from the j-th niche at time t, and is the affinity- and abundance-weighted proportion of the total rate p_j at which

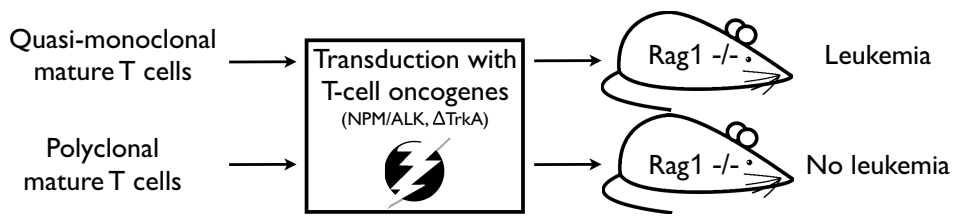


FIGURE 1 | Experimental strategy as described in Newrzela et al. (21). TCR quasi-monoclonal T-cell populations transduced with potent T-cell oncogenes developed mature T-cell lymphoma/leukemia in RAG1-deficient recipient mice, while TCR polyclonal T-cell populations that have been subjected to the same transformation assay did not.

niche j provides resources. Thus, $r_{ij}(t)$ is calculated as follows:

$$r_{ij}(t) = \begin{cases} \frac{a_{ij}c_i(t)}{\sum_{k=1}^{m} a_{kj}c_k(t)} p_j & \text{if } \sum_{k=1}^{m} a_{kj}c_k(t) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The sum of the resource acquisition rates over all niches yields the net resource acquisition rate for a particular species.

Different T-cell species may differ in the amount of resources needed to sustain an individual cell. Therefore, we introduce the m-dimensional vector $\mathbf{v} = (v_i)$, which we term *resource utilization efficiency*. This parameter determines for each species, how many T cells can be sustained by one resource unit per unit of time. The product of the resource acquisition rates and the resource utilization efficiencies yields the time-dependent carrying capacities of the individual species, stored in the vector $\mathbf{k}(t) = (k_i(t))$:

$$k_i(t) = v_i \sum_{j=1}^{n} r_{ij}(t)$$

The carrying capacities indicate how many cells of a given species can be sustained by the system, given the current resource distribution among the species. The carrying capacities may change dynamically in time, as the species abundances vary. The relation between the actual size of a species and the carrying capacity determines whether the species abundance decreases or increases. Due to lack of biological data determining a particular growth model for T-cell clones, we chose a rather general logistic growth dynamics, where τ denotes the minimum cell cycle time:

$$\frac{dc_i(t)}{dt} = \begin{cases} \frac{c_i(t)}{\tau}\left(1 - \frac{c_i(t)}{k_i(t)}\right) & \text{if } k_i(t) > 0 \\ 0 & \text{if } k_i(t) = 0 \end{cases}$$

We require c_i^0, a_{ij}, p_j, and v_i to be greater than or equal to zero, and τ to be greater than zero.

2.2. PARAMETER CHOICE

In its general form, the model presented in the previous subsection has $nm + 2m + n + 1$ scalar parameters. In the following, we sketch our numerical approach and reparameterize the model in order to reduce the effective number of parameters.

In our approach, the number of TCR-defined clones is denoted by q. Each clone is represented by two species (thus $m = 2q$). Without loss of generality, species 1 to q represent the presumably healthy cells within the q clones (referred to as *healthy species*) and species $q+1$ to $2q$ represent the cells that potentially give rise to leukemia/lymphoma (referred to as *preleukemic species*). The specific case of the monoclonal situation is represented by setting all but the first healthy and the first preleukemic initial species abundances to zero ($c_i = 0$ except $c_1 > 0$, $c_{q+1} > 0$). We assume that each species may in principle receive a stimulus from each niche due to unspecific affinity. The magnitude of the unspecific affinity is denoted $u^{(h)}$ for the healthy species and $u^{(p)}$ for the preleukemic species. In addition, we assume that each species has a preferred niche, to which it has an additional affinity [*specific affinity*, denoted $s^{(h)}$ for the healthy species and $s^{(p)}$ for the preleukemic species]. Biologically, the specific affinity could be interpreted as affinity to cognate self-peptide, and the unspecific affinity as affinity to the MHC itself.

The scalar parameters describing the specific and the unspecific affinities are used to define the niche affinity matrix $\mathbf{A}$. Formally, we construct the matrices $\mathbf{A}^{(h)} = \left[a_{ij}^{(h)}\right]$ describing the niche affinities of the healthy species and $\mathbf{A}^{(p)} = \left[a_{ij}^{(p)}\right]$ describing the niche affinities of the preleukemic species as follows,

$$a_{ij}^{(h)} = \begin{cases} s^{(h)} + u^{(h)} & \text{if } i = j \\ u^{(h)} & \text{if } i \neq j \end{cases}$$

$$a_{ij}^{(p)} = \begin{cases} s^{(p)} + u^{(p)} & \text{if } i = j \\ u^{(p)} & \text{if } i \neq j \end{cases}$$

$$i \in \{1, 2, ..., q\}, \qquad j \in \{1, 2, ..., n\}$$

and use these matrices in order to construct the actual niche affinity matrix $\mathbf{A}$ by vertical concatenation of $\mathbf{A}^{(h)}$ and $\mathbf{A}^{(p)}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{(h)} \\ \mathbf{A}^{(p)} \end{bmatrix}$$

Furthermore, we introduce the symbols $v^{(h)}$ and $v^{(p)}$ to describe the resource utilization efficiencies of the healthy and preleukemic species, respectively. $v^{(h)}$ and $v^{(p)}$ are used to construct the actual resource utilization efficiencies (superscript T denoting the transpose of a vector):

$$\mathbf{v} = (\underbrace{v^{(h)}, \ldots, v^{(h)}}_{q \text{ times}}, \underbrace{v^{(p)}, \ldots, v^{(p)}}_{q \text{ times}})^T$$

Note that no interspecies heterogeneity with respect to the magnitudes of the specific and unspecific affinities of the healthy species ($s^{(h)}$, $u^{(h)}$) and their resource utilization efficiency $v^{(h)}$ is considered. The same applies to the preleukemic species, with parameters $s^{(p)}$, $u^{(p)}$, and $v^{(p)}$.

In order to make numerical simulations feasible, while still preserving the niche-based regulation, we are using a system with 100 niches ($n = 100$) and 100 clones ($q = 100$), i.e. 200 species ($m = 200$) for the numerical simulations.

As mentioned above, the real number of T-cell clones in mice is estimated in the order of 10^6, while the total number of T cells in a mouse is estimated to be in the order of 10^8 (23). Assuming that in the physiological situation the number of niches in mice equals the number of T-cell clones, a niche nourishes 100 cells on average. Therefore, we consider the niche sizes $p_j = 100$ for all $j = 1, \ldots, 100$. The minimum cell cycle time τ is set to 8 h as a rough estimate of vigorous T-cell proliferation (24). Since an initial exploratory screen did not reveal a qualitative effect of changes in niche size $\mathbf{p}$ and the minimum cell cycle time τ on the model behavior, we keep $\mathbf{p}$ and τ fixed to the above described values.

Additionally, and without loss of generality, we can fix one of the four niche affinity parameters $s^{(h)}$, $s^{(p)}$, $u^{(h)}$, and $u^{(p)}$ as the distribution of resources depends on relative affinities only. We chose to fix the unspecific affinity of the healthy cells $u^{(h)} = 1/n = 0.01$

(n being the number of niches) as our reference. The specific affinity of the healthy species $s^{(h)}$ is set to 1 [except for the results shown in **Figure 6**, in which we are also considering the values $s^{(h)} = 0.2$ and $s^{(h)} = 5$]. Furthermore, we set the resource utilization efficiency of the healthy cells $v^{(h)}$ to 1, thus interpreting the corresponding value of the preleukemic cells $v^{(p)}$ as a relative measure compared to the healthy situation.

Hence, at this point the scalar parameters $s^{(p)}$ (specific affinity of the preleukemic species), $u^{(p)}$ (unspecific affinity of the preleukemic species), $v^{(p)}$ (resource utilization efficiency of the preleukemic species), and the vector-valued initial condition $\mathbf{c}^0$ are not yet fixed. How they are varied in the parameter screen is described in the following subsection.

2.3. SIMULATION PROCEDURE

2.3.1. Physiological situation

In order to study the behavior of the system, we *in silico* transplant 500 cells into an empty system. TCR diversities $q = 1$ and $q = 100$ are considered. In the TCR polyclonal situation, the initial species abundances $\mathbf{c}^0$ are obtained by distributing the 500 cells according to a uniform distribution over the healthy species. $s^{(h)}$ is set to 1. The parameters $s^{(p)}$, $u^{(p)}$, and $v^{(p)}$, which describe the properties of the preleukemic species, do not influence the model behavior here, since no preleukemic cells are transplanted into the system. All other model parameters are fixed to the above described values. We let the system evolve for 400 time-steps. To demonstrate the system response to perturbations, we simulate a significant cell loss by removing 99% of the cells at time $t = 200$. The simulation results are presented in subsection 1.

2.3.2. Oncogenic situation

In order to systematically evaluate the effect of the parameters describing the properties of the preleukemic cells [i.e. their specific affinity $s^{(p)}$, their unspecific affinity $u^{(p)}$, their resource utilization efficiency $v^{(p)}$] and the initial abundances $\mathbf{c}^0$, we perform an extensive parameter screen. $s^{(p)}$, $u^{(p)}$, and $v^{(p)}$ are varied in multiples of their healthy counterparts, ranging from $\sim$1/40 to $\sim$40. The relative distances between two neighboring values is 20%, so that 41 different fold-changes are considered per parameter.

The fraction of transplanted cells that have been transformed into preleukemic cells is currently not known for the used experimental protocol. Therefore, we consider three scenarios, denoted P1, P10, and P100, which consider 1, 10, and 100 initial preleukemic cells, respectively. All three scenarios are evaluated for each combination of the specific affinity of the preleukemic species $s^{(p)}$, their unspecific affinity $u^{(p)}$, and their resource utilization efficiency $v^{(p)}$. All cases are initiated by 500 cells that are *in silico-transplanted* into an empty system.

This corresponds to 5 cells per clone in the polyclonal situation on average, in accordance with the experimental situation (5×10^6 transplanted cells, TCR diversity estimated in the order of 10^6). In scenario P1, one T cell (0.2%) is assigned to the preleukemic cell compartment, in scenario P10, 10 T cells (2%), belonging to 10 different species in the TCR polyclonal situation, are assigned to the preleukemic cell compartment, and in scenario P100, 100 T cells (20%), belonging to 100 different species in the polyclonal situation. In all three scenarios, the remaining cells are distributed randomly according to a uniform probability distribution over the healthy species in the polyclonal situation. In addition to the polyclonal settings we construct a corresponding monoclonal situation, in which the fraction of 0.2% (P1), 2% (P10), or 20% (P100) preleukemic cells is assigned to only a single clone.

The number of parameter sets evaluated in each of the three scenarios is 41^3 (since 41 different fold-changes are considered for $s^{(p)}$, $u^{(p)}$, and $v^{(p)}$). Hence, in total $3 \times 41^3 \approx 2 \times 10^5$ parameter sets are evaluated for both the polyclonal and the monoclonal situation.

As we are not primarily interested in transient phenomena but in the long-term behavior of the system, our aim is the identification and characterization of stable steady states, namely whether the preleukemic cells are able to dominate the system or not. Therefore, simulations are run until the relative change of all species abundances between two successive time steps are below 10^{-6}. For this situation we assume that the system is sufficiently close to a steady state. If this criterion is not fulfilled after 10^8 simulation steps, the current simulation is stopped, and the stability of the system will be assessed manually.

For each individual parameter set, we evaluate if the corresponding simulations for the mono- and polyclonal situation are consistent with the experimental phenomena, i.e. if we observe a stable and considerably enlarged population of preleukemic cells in the monoclonal situation, and control of the preleukemic cells in the polyclonal situation (i.e. the contribution of the preleukemic cell population remains below a certain threshold). The specific criteria used for the classification are listed in **Table 1**. If a parameter set fulfills all criteria, we consider it consistent with the observed phenomena as described in (20, 21).

The simulation results of the parameter screen are presented in subsection 2.

3. RESULTS

3.1. PHYSIOLOGICAL SITUATION

First, we investigate the behavior of the described system in the absence of preleukemic cells. As can be seen in **Figure 2**, the system quickly converges to a steady state, both in the mono- and the

Table 1 | Criteria used for classification.

	Total cell count	Contribution of preleukemic cells
TCR monoclonal situation	At least 300% of physiological cell count	At least 80% of total cell count
TCR polyclonal situation	At most 120% of physiological cell count	At most 50% of total cell count

A particular tested parameter set is classified as consistent with the experimental phenomena, if the total cell count and the contribution of the preleukemic cells fulfill the criteria specified in the table, i.e. a leukemia-like situation in the TCR monoclonal situation, and a state of non-dominance of the preleukemic species in the TCR polyclonal situation. The total cell count at the steady state is compared to the physiological cell count, which is established based on the simulations shown in subsection 1. The contribution of the preleukemic cells is the ratio between the number of preleukemic cells and the total cell count at the steady state.

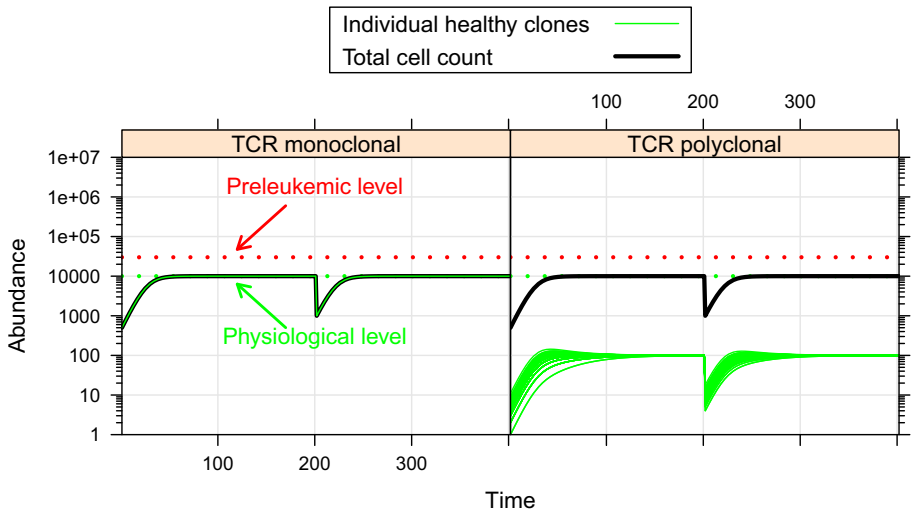


FIGURE 2 | Simulation results of physiological situation. The left panel documents the dynamic system behavior in the monoclonal situation. In the right panel, the individual green lines represent the abundance of different clones. The system quickly converges to a steady state. The total cell count in the steady state of the physiological situation is indicated by the horizontal dashed green line at 10^4. The horizontal red line at 3×10^4 represents the total cell count that is required as a minimum in order to qualify a situation as premalignant. At $t = 200$, 99% of the cells are removed from the system. Both in the mono- and the polyclonal situation the system quickly reestablishes the previous steady state.

polyclonal situation. In the polyclonal situation, all clones have an abundance of 100 cells at the steady state due to the parameter symmetry among the healthy species (i.e. all species have the same specific affinity $s^{(h)} = 1$ to their preferred niche, and the same unspecific affinity $u^{(h)} = 1/n$ to all niches). The total cell count at the steady state amounts to $v^{(h)} \cdot \sum_{j=1}^{n} p_j = 10^4$ both in the monoclonal and the polyclonal situation. After a perturbation, e.g. due to cell kill, the system quickly reestablishes its previous state, unless individual species are completely eliminated within the cell kill simulation.

The steady state with equal-sized clones is reached regardless of the initial abundances of the healthy cells, given that the abundances of all healthy species are >0 (data not shown). Transient changes of the niche sizes or transient addition of niches (e.g. to model infections) can entail the transient expansion of one more species/clones (data not shown).

3.2. ONCOGENIC SITUATION

For the oncogenic situation, we evaluate whether a certain proportion of preleukemic cells (represented by the three scenarios P1, P10, and P100) can develop into a (pre)leukemic situation in the monoclonal case, while being controlled by healthy competitor cells in the polyclonal case. The evaluation is carried out based on prespecified formal criteria (see **Table 1**). In all tested parameter settings, the steady state (i.e. convergence) criterion is reached both in the mono- and the polyclonal situation. **Figure 3** shows representative simulation results to provide some intuition about the spectrum of possible simulation outcomes for the three scenarios P1, P10, and P100.

Out of the $\approx 2 \times 10^5$ tested parameter configurations, 1050 parameter sets are consistent with the experimental phenomena according to the defined criteria (c.f. **Table 1**). Further, it should be emphasized that the consistent parameter sets are identifiable as a confined region in the parameter space in all three scenarios (see **Figure 4**) and that the regions in scenarios P1, P10, and P100 overlap considerably. In all three scenarios, we identify a region that is characterized by lowered specific and unspecific affinities as well as an increased resources utilization of preleukemic compared to normal T cells. Technically, this refers to triangular prisms in the $u_{low}^{(p)} s_{low}^{(p)} v_{high}^{(p)}$ octant of the parameter cube (**Figure 4**). In scenario P1, the region of consistent parameter sets additionally extends to the $u_{low}^{(p)} s_{high}^{(p)} v_{high}^{(p)}$ octant. Our results indicate that the overall systems behavior displays only a minor dependency on the initial number of preleukemic cells with respect to the phenomena in focus.

Assessing the set of consistent parameter constellations in more detail, we find that the resource utilization efficiency parameter for the preleukemic species $v^{(p)}$ had to be chosen at least three-fold higher than the resource utilization efficiency for healthy cells, in order to be consistent with the experimental observations. This is more clearly seen in **Figure 5**, which further characterizes the parameter sets that are consistent with the predefined criteria. The fact that a three-fold increase of $v^{(p)}$ is required, directly reflects criterion from **Table 1**, stating that the cell counts have to be increased at least three-fold in the monoclonal situation.

Concerning the niche affinities of the preleukemic cells, a clear pattern can be observed, likewise apparent in **Figure 5**. In scenarios P10 and P100, the specific affinity of the preleukemic cells $s^{(p)}$ is decreased (i.e. $s^{(p)} < s^{(h)}$) in all consistent parameter sets. In scenario P1, however, the specific affinity can be considerably increased without rendering a tested parameter set inconsistent. This is due to the fact that in scenario P1, only a single cell is preleukemic initially. Therefore, the fitness of the preleukemic cell

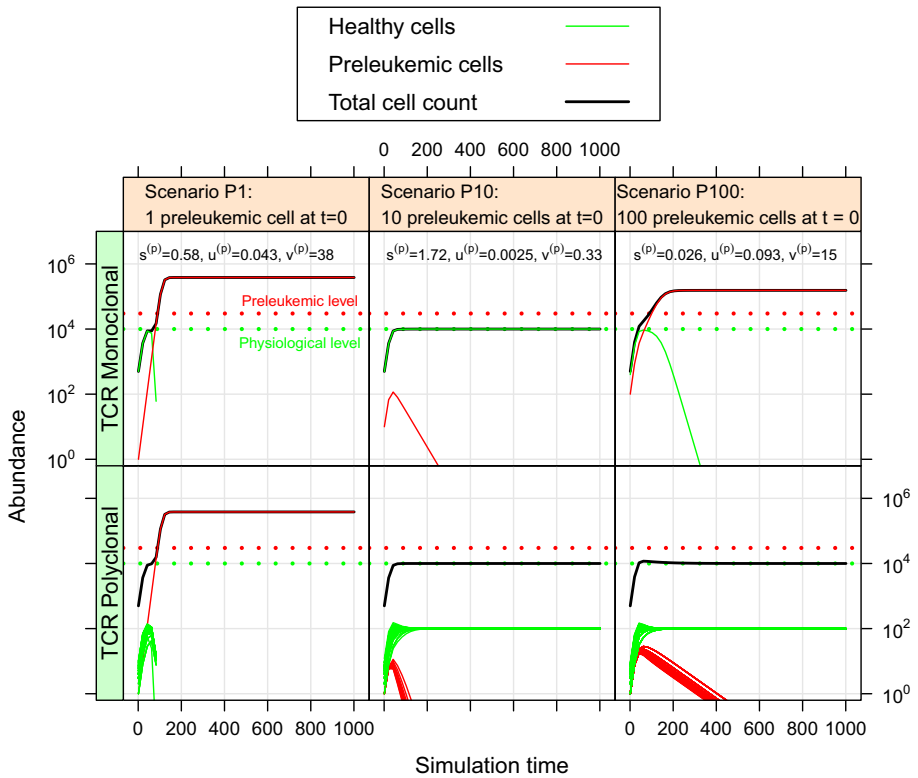


FIGURE 3 | Simulation results of three representative parameter sets. For each scenario, we show the simulation results of one possible parameter set. The parameter set picked for scenario P1 is not consistent with the experimental observations as specified in **Table 1**. The preleukemic cells expand more than threefold both in the mono- and polyclonal situation. The parameter set chosen for P10 is also not consistent with the specified criteria. Using this parameter set, the preleukemic cells are less fit than the healthy cells, and die out both in the mono- and the polyclonal situation. The parameter set chosen for scenario P100 is consistent with the specified criteria.

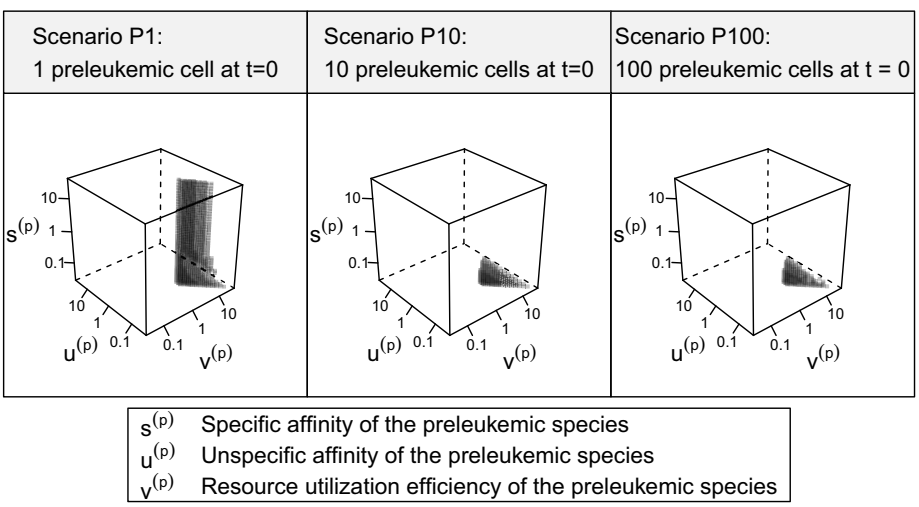


FIGURE 4 | 3D scatter plot of consistent parameter sets. Each tested parameter set can be represented by a point in one of the three presented cubes, the coordinates representing the fold-changes of $s^{(p)}$, $t^{(p)}$, and $v^{(p)}$ in comparison with their healthy counterparts $s^{(h)}$, $u^{(h)}$, and $v^{(h)}$. In the center of each cube, the parameters $s^{(p)}$, $u^{(p)}$, and $v^{(p)}$ are equal to their counterparts describing the healthy cells. Only the parameter sets that are consistent with the experimental observations (criteria in **Table 1**) are plotted. Transparency is used in order to give an impression of the shape of the consistent parameter region.

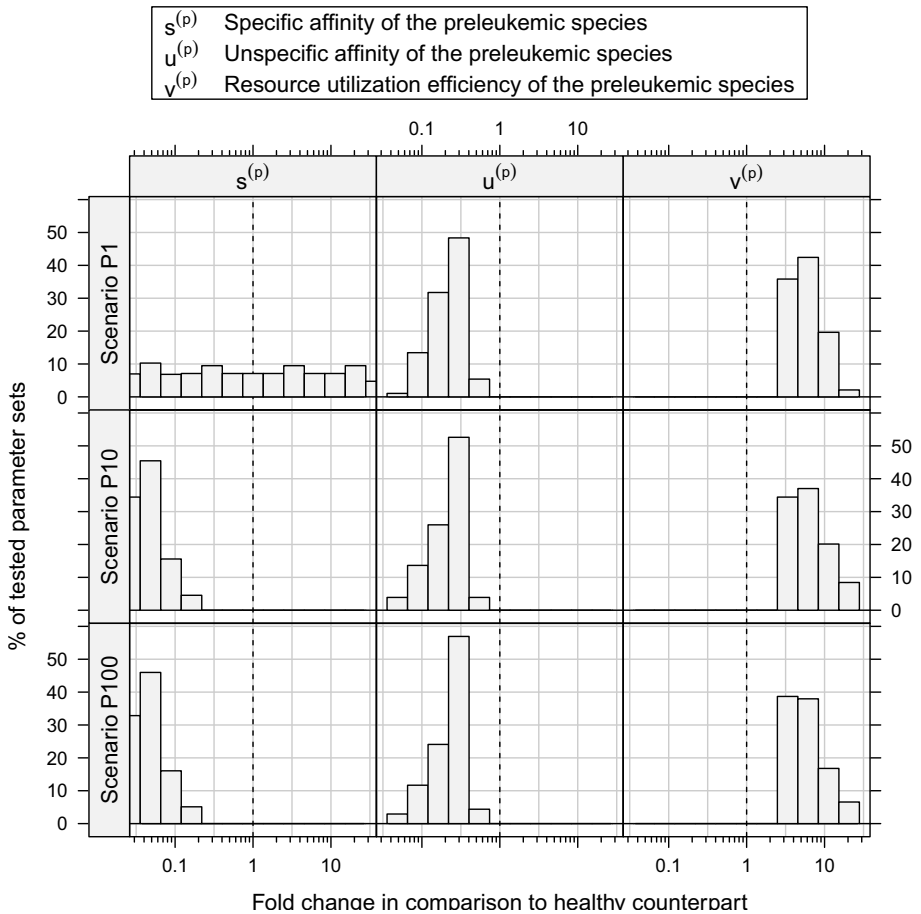


FIGURE 5 | Histograms of consistent parameter values. These histograms show the distribution of the specific affinity of the preleukemic species $s^{(p)}$, their unspecific affinity $u^{(p)}$, and their resource utilization efficiency $v^{(p)}$ of the consistent parameter sets. $s^{(p)}$ is decreased in all consistent parameter sets of scenario P10 and P100, while it is indifferent in scenario P1. $u^{(p)}$ is decreased and $v^{(p)}$ is increased in all consistent parameter sets.

pool (i.e. proliferative potential) increases only marginally in the polyclonal situation if the specific affinity of the preleukemic cells $s^{(p)}$ is increased. In contrast, in scenario P10 and P100, the fitness of the preleukemic cell population in the polyclonal situation responds more sensitively to an increase of the specific affinity of the preleukemic cells due the greater TCR diversity within the preleukemic cell population. Because of that, the preleukemic cells cannot be controlled by healthy competitor cells in the polyclonal situation, even if the specific affinity is only mildly increased, rendering such parameter sets inconsistent.

The unspecific affinity of the preleukemic cells is decreased in all consistent parameter sets (i.e. $u^{(p)} < u^{(h)}$). This reduction generates a disadvantage of preleukemic species for accessing resources from non-preferred niches, thus prohibiting a dominating expansion. Nonetheless, the preleukemic species must have at least some residual unspecific affinity in order to generate results that are consistent with the experimental situation. Specifically, the residual unspecific affinity allows them to access resources from the unprotected niches in the monoclonal situation. Without this ability, these cells cannot outgrow the healthy species in this situation. When comparing the decrease of the specific and the unspecific affinity of the preleukemic cells ($u^{(p)}$), it stands out that in scenarios P10 and P100, the specific affinity of the preleukemic cells ($s^{(p)}$) is decreased to a greater degree than the unspecific affinity ($u^{(p)}$) in all consistent parameter sets. In contrast, the decrease of the specific affinity of preleukemic cells ($s^{(p)}$) is greater than their decrease in unspecific affinity ($u^{(p)}$) in only ≈30% of the consistent parameter sets in scenario P1. Hence, if only a single preleukemic cell is present initially, the ability to receive a stimulus due to interaction with cognate self-peptide may be preserved or even increased, while in the situation of many initially present preleukemic cells the simulation results are not consistent with the experimental phenomena.

Furthermore, our simulations demonstrate that there is a rather strict functional relation between the unspecific affinity of the preleukemic cells ($u^{(p)}$) and their resource utilization efficiency ($v^{(p)}$) for all consistent parameters (see **Figure 6**). In other words, a certain acquired growth advantage, which is mediated by an increase in resource utilization ($v^{(p)}$), requires a corresponding decrease in the unspecific affinity ($u^{(p)}$) to guarantee consistency. Interestingly, the stringency of this relation depends crucially on the specific affinity of the healthy cells ($s^{(h)}$). To illustrate this

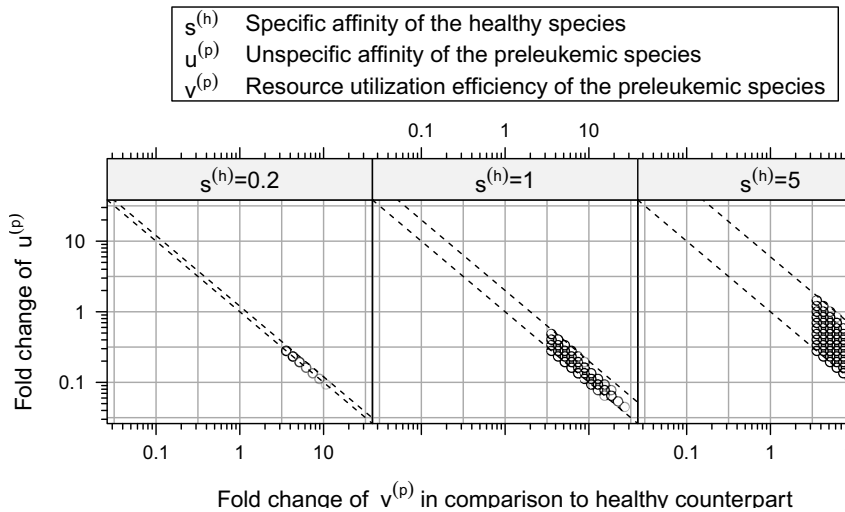


FIGURE 6 | Relation between resource utilization efficiency $v^{(p)}$ and unspecific affinity $u^{(p)}$ in the consistent parameter sets. Here, the fold-changes of the unspecific affinities of the preleukemic cells $u^{(t)}$ in the consistent parameter sets is plotted against their resource utilization efficiencies $v^{(t)}$. The relation between these quantities is relatively strict. All points are contained between two parallel lines. In this plot, we show additional simulation results, in which also the specific affinity of the healthy cells $s^{(h)}$ is varied.

relationship, we tested the system behavior also for $s^{(h)} = 0.2$ and $s^{(h)} = 5$ in addition to the value $s^{(h)} = 1$ used so far. Although varying values of $s^{(h)}$ do not qualitatively change the system behavior, the size of the consistent parameter region, i.e. the number of consistent parameter sets, is affected considerably by the particular choice of $s^{(h)}$: it is larger for $s^{(h)} = 5$ and smaller for $s^{(h)} = 0.2$. This can be seen in **Figure 6**, which illustrates the relationship for different values of $s^{(h)}$.

Interpreting the specific affinity as affinity to cognate self-peptide, and the unspecific affinity as affinity to the MHC itself, the ratio of the specific and the unspecific affinity of the healthy cells (i.e. $s^{(h)}/u^{(h)}$) determines how strict the regulation exerted by the self-peptide defined niches is. If the ratio is large, the niche regulation is strict, i.e. T cells can hardly access resources from niches other than their preferred niche. If the ratio is small, T cells can acquire more resources from niches other than their preferred niche. Our simulation results indicate that the stronger the niche regulation (i.e. a larger $s^{(h)}$), the larger is the consistent parameter region, and hence the more robust are the modeling results and the biological behavior of the system.

4. DISCUSSION

Oncogenesis in mature T cells and especially the resistance to malignant transformation of these cells even in potent oncogenic transformation assays have raised substantial interest in these cell populations. In a series of experiments, it has been shown that the resistance of mature T cells depends on the diversity of the TCR repertoire. Specifically, MTCLLs could readily be induced in T-cell populations that are quasi-monoclonal with respect to their TCR, while oncogenesis occurred only in rare cases in the TCR polyclonal situation, despite identical experimental conditions (20, 21). This observation led to the hypothesis that in the polyclonal situation, regulation based on competition for survival stimuli mediated by interaction between TCRs and spMHCs on APCs can prevent the expansion of preleukemic cells. Furthermore, the hypothesis implies that this regulation fails in the quasi-monoclonal situation, and the preleukemic cells expand, and eventually give rise to overt malignancy.

We have developed and applied a mathematical model of T-cell homeostasis to qualitatively and quantitatively challenge this concept. In the model, mature T cells compete for resources that are provided by T-cell niches. Biologically, the resources correspond to survival stimuli from APCs due to interactions between spMHCs and the TCR. The different T-cell niches are defined by the different spMHCs found on APCs. Due to differential affinities of different TCRs to spMHCs, T cells may differ in their capability to acquire resources from a particular niche.

Within this modeling framework we systematically evaluated the effect of the model parameters. For each specific parameter set, we tested if it is consistent with the observed phenomena according to the criteria in **Table 1**.

In order to achieve consistency with the criteria, the niche affinities of the preleukemic cells had to be decreased (i.e. $s^{(p)} < s^{(h)}$ and $u^{(p)} < u^{(h)}$, while their resource utilization efficiency had to be increased (i.e. $v^{(p)} > v^{(h)}$). Biologically, this suggests that the preleukemic cells are less dependent on niche resources (decreased affinities). In particular, our results predict a stronger decrease of the specific affinity of the preleukemic cells ($s^{(p)}$) compared to the unspecific affinity ($u^{(p)}$).

This implies that in the presence of healthy competitor cells with preference for the same niche, the preleukemic cells can be outcompeted. In the polyclonal situation this preventive effect acts on almost all niches, thus keeping the preleukemic cells at a minimum or even making them disappear.

In contrast, in the monoclonal situation the vast majority of niches is unguarded by healthy cells, and their resources can,

therefore, be accessed by the preleukemic cells due to their albeit small, but non-zero unspecific affinity ($u^{(p)}$). In this situation, the decrease in the affinities can be compensated by the increase of the resource utilization efficiency of the preleukemic cells ($v^{(p)}$). Metaphorically speaking, the polyclonal healthy competitor cells protect the niche resources against the aggressive preleukemic cells in the polyclonal situation. If the polyclonal guards in the form of healthy competitor cells are not present, the preleukemic cells are able to claim all or most of the niche resources and promote MTCLL occurrence.

Transferring our findings back into a mechanistic context, we speculate that the decreased model affinities correspond to a down-regulation of TCR-mediated regulations and/or impairment of intracellular TCR signaling, i.e. a reduced T-cell avidity (25). A more efficient utilization of these signals (i.e. an increase in $v^{(p)}$) can be interpreted as a partial independence of preleukemic cells from environment-mediated growth signals, or as failure to respond adequately to inhibitory signals. On an intracellular level this independence might be achieved by a constitutive up-regulation of growth-promoting pathways.

In order to describe the T-cell system, and in particular the experiments performed by Newrzela et al. in mathematical terms we made a number of simplifying assumptions. It is a long-standing notion that oncogenesis often develops in a multi-step process (26). Also, in the context of gene therapy, it has been described that oncogenesis may occur directly (*single hit induction*), or may require additional lesions (*cooperating hit induction*) (27, 28). In our simulation scenario, however, we consider the effect of a first but fully effective hit only, i.e. we assume a virtually instantaneous effect of the oncogene affecting all preleukemic cells after transduction. In principle, it would be possible to incorporate cooperating hits into our model that are subsequently acquired after transplantation. We refrained from doing so, since only preliminary experimental data on tumor evolution in the context of the phenomena in focus are available to date.

Similarly, we did not account for heterogeneity concerning the effect of the oncogene, i.e. all cells belonging to the preleukemic cell compartment are assigned the same functional alteration in the model. Also, it is not clear, if our model adequately describes the emergence of preleukemic cells on an individual clone level. Hopefully, such individual clone data will be available in the future, so that we will be able to validate our model in this regard.

In order to make numeric simulations feasible, we considered a system with only 100 niches and 100 T-cell clones in the polyclonal situation. This is several orders of magnitude below the estimated TCR diversity in mice, and the estimated number of self-peptides that can be presented per allele [$\sim 10^5$ Ref. (13)]. However, there is significant cross-reactivity (i.e. presence of TCRs that may recognize more than one spMHC) in the T-cell system (29), so that the number of functionally relevant, distinct niches may be considerably lower. Being well aware of these simplifications, we consider the proposed system dimensions as sufficient in order to capture the general niche-based structure of the mature T-cell system.

Although we successfully demonstrated that suppression of preleukemic T cells due to TCR-related regulation in the polyclonal situation can consistently explain the experimental observations, we are aware that alternative explanations could also account for the observed phenomena. Definitive proof or disproof of the niche regulation hypothesis will require further experimental work.

Therefore, our conceptual understanding of T-cell oncogenesis will be challenged in further experiments. Specifically, we suggest to compare the phenotype of the leukemic/preleukemic cells with the phenotype of the healthy competitors regarding, e.g. their gene expression profile, their activation of relevant signaling pathways and their functional properties (30). This will allow assessing whether our assertions about the properties of the preleukemic cells (i.e. down-regulation of the TCR/TCR signaling and up-regulation of growth-promoting pathways) are correct. Our modeling results predict that there is a minimum clonal diversity that is needed in order to control preleukemic T cells as produced in the previous experiments. Further transplantation experiments are planned to determine the minimum TCR diversity in clonality titration experiments.

All simulations presented in this publication start with a hypocellular condition. However, additional simulations (data not shown) demonstrated that the model results are generally not dependent on the initial abundance of a species, as long as it is present at all. Therefore, our results apply also to leukemogenesis with physiological onset conditions as it occurs in clinical settings. The situation in which the malignancy develops from a single mutated cell (as can be expected for the majority of patients) corresponds to the presented scenario P1.

Central to this work is our assumption that the abundances of individual TCR-defined clones are regulated on a per-clone basis due to interactions with spMHCs. However, many aspects of the insinuated regulation are elusive to date. This is at least partially due to the fact that in the *in vivo* situation, accurate and precise time-dependent quantification of individual TCR-defined clones with abundances in the physiological range is technologically challenging, if not impossible, to date. Nonetheless, such a concept is intellectually appealing, and seems plausible on a mechanistic level. We hypothesize that this regulation is (at least in part) responsible for the relatively low incidence of mature T-cell malignancies relative to mature B-cell malignancies. This hypothesis would imply less effective control mechanisms in the mature B-cell system, which might be caused by the fact that clonal homeostasis in the mature B-cell system is presumably more complex than in the T-cell system, due to affinity maturation of the B-cell receptor, and ongoing influx of newly generated mature B-cell clones from the bone marrow. These complicating factors may undermine the effectiveness of the leukemia control mechanisms proposed for the mature T-cell system in the mature B-cell system.

So far in this publication, the unspecific affinities $u^{(h)}$ and $u^{(p)}$ are interpreted in terms of TCR affinity for the MHC molecule itself. However, these unspecific affinities could also represent the reliance of T-cell survival on survival cytokines, e.g. interleukin 7. Further studies could aim to disentangle these two potential components of the unspecific affinity. To do so, the use of mouse models with a restricted MHC repertoire might be useful.

As exemplified in this paper, mathematical modeling approaches allow for the quantitative assessment of functional principles of T-cell interactions, their integration into a consistent

conceptual framework and the derivation of testable hypotheses. Our results add a new puzzle piece to the complex picture of mature T-cell homeostasis. The hypothesis that regulation based on TCR-defined clonal membership suppresses MTCLL emergence may serve as a novel starting point to delve deeper into the mechanisms governing the homeostatic behavior of mature T cells.

Our modeling results prompt further experimental research to clarify the nature of the differential transformability of mature T cells. Moreover, our work demonstrates the general ability of theoretical approaches to formalize and conceptually validate the results of experimental research and promotes the idea of an iterative, interdisciplinary approach to research in immunology.

5. AUTHORS CONTRIBUTION

Sebastian Gerdes performed mathematical modeling and wrote the paper, Sebastian Newrzela provided conceptual input and wrote the paper, Ingmar Glauche performed mathematical modeling and wrote the paper, Martin-Leo Hansmann provided conceptual input, Dorothee von Laer provided conceptual input, Ingo Roeder wrote the paper, provided conceptual input, and guided research.

ACKNOWLEDGMENTS

Ingo Roeder, Sebastian Newrzela, and Martin-Leo Hansmann are funded by the German Research Foundation (DFG) under RO3500/4-1, NE1438/4-1, and HA1284/8-1 as part of the collaborative research group on mature T-cell lymphomas, "CONTROL-T." Furthermore, Sebastian Gerdes and Ingo Roeder were funded by the DFG grant RO3500/1-2.

REFERENCES

1. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size estimate of the αßTCR repertoire of naive mouse splenocytes. *J Immunol* (2000) **164**(11):5782–7.
2. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, et al. The influence of age on T cell generation and TCR diversity. *J Immunol* (2005) **174**(11):7446–52.
3. Jung D, Alt FW. Unraveling V(D)J recombination: insights into gene regulation. *Cell* (2004) **116**(2):299–311. doi:10.1016/S0092-8674(04)00039-X
4. Huesmann M, Scott B, Kisielow P, von Boehmer H. Kinetics and efficacy of positive selection in the thymus of normal and T cell receptor transgenic mice. *Cell* (1991) **66**(3):533–40. doi:10.1016/0092-8674(81)90016-7
5. Minter LM, Osborne BA. Cell death in the thymus – it' s all a matter of contacts. *Semin Immunol* (2003) **15**(3):135–44. doi:10.1016/S1044-5323(03)00029-0
6. Zal T, Volkmann A, Stockinger B. Mechanisms of tolerance induction in major histocompatibility complex class II-restricted T cells specific for a blood-borne self-antigen. *J Exp Med* (1994) **180**(6):2089–99. doi:10.1084/jem.180.6.2089
7. Freitas AA, Rocha B. Population biology of lymphocytes: the flight for survival. *Annu Rev Immunol* (2000) **18**(1):83–111. doi:10.1146/annurev.immunol.18.1.83
8. Singh NJ, Bando JK, Schwartz RH. Subsets of nonclonal neighboring CD4+ T cells specifically regulate the frequency of individual antigen-reactive T cells. *Immunity* (2012) **37**(4):735–46. doi:10.1016/j.immuni.2012.08.008
9. Brocker T. Survival of mature CD4 T lymphocytes is dependent on major histocompatibility complex class II-expressing dendritic cells. *J Exp Med* (1997) **186**(8):1223–32. doi:10.1084/jem.186.8.1223
10. Kirberg J, Berns A, von Boehmer H. Peripheral T cell survival requires continual ligation of the T cell receptor to major histocompatibility complex-encoded molecules. *J Exp Med* (1997) **186**(8):1269–75. doi:10.1084/jem.186.8.1269
11. Martin B, Bécourt C, Bienvenu B, Lucas B. Self-recognition is crucial for maintaining the peripheral CD4+ T-cell pool in a nonlymphopenic environment. *Blood* (2006) **108**(1):270–7. doi:10.1182/blood-2006-01-0017
12. Hataye J, Moon JJ, Khoruts A, Reilly C, Jenkins MK. Naive and memory CD4+ T cell survival controlled by clonal abundance. *Science* (2006) **312**(5770):114–6. doi:10.1126/science.1124228
13. Mahajan VS, Leskov IB, Chen JZ. Homeostasis of T cell diversity. *Cell Mol Immunol* (2005) **2**(1):1–10.
14. Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD, Linet MS. Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. *Blood* (2006) **107**(1):265–76. doi:10.1182/blood-2005-06-2508
15. Deeks S, Wagner B, Anton P, Mitsuyasu R, Scadden D, Haung C, et al. A phase II randomized study of HIV-specific T-cell gene therapy in subjects with undetectable plasma viremia on combination antiretroviral therapy. *Mol Ther* (2002) **5**(6):788–97. doi:10.1006/mthe.2002.0611
16. Recchia A, Bonini C, Magnani Z, Urbinati F, Sartori D, Muraro S, et al. Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. *Proc Natl Acad Sci U S A* (2006) **103**(5):1457–62. doi:10.1073/pnas.0507496103
17. Scholler J, Brady TL, Binder-Scholl G, Hwang W-T, Plesa G, Hege KM, et al. Decade-long safety and function of retroviral-modified chimeric antigen receptor T cells. *Sci Transl Med* (2012) **4**(132):ra53–132. doi:10.1126/scitranslmed.3003761
18. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack M, Wulffraat N, Leboulch P, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* (2003) **302**(5644):415–9. doi:10.1126/science.1088547
19. Ott M, Schmidt M, Schwarzwaelder K, Stein S, Siler U, Koehl U, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat Med* (2006) **12**(4):401–9. doi:10.1038/nm1393
20. Newrzela S, Cornils K, Li Z, Baum C, Brugman MH, Hartmann M, et al. Resistance of mature T cells to oncogene transformation. *Blood* (2008) **112**(6):2278–86. doi:10.1182/blood-2007-12-128751
21. Newrzela S, Al-Ghaili N, Heinrich T, Petkova M, Hartmann S, Rengstl B, et al. T-cell receptor diversity prevents T-cell lymphoma development. *Leukemia* (2012) **26**(12):2499–507. doi:10.1038/leu.2012.142
22. De Boer RJ, Perelson AS. T cell repertoires and competitive exclusion. *J Theor Biol* (1994) **169**(4):375–90. doi:10.1006/jtbi.1994.1160
23. Doherty P, Riberdy J, Belz G. Quantitative analysis of the CD8+ T-cell response to readily eliminated and persistent viruses. *Philos Trans R Soc Lond B Biol Sci* (2000) **355**(1400):1093–101. doi:10.1098/rstb.2000.0647
24. De Boer RJ, Perelson AS, Ribeiro RM. Modelling deuterium labelling of lymphocytes with temporal and/or kinetic heterogeneity. *J R Soc Interface* (2012) **9**(74):2191–200. doi:10.1098/rsif.2012.0149
25. McKee MD, Roszkowski JJ, Nishimura MI. T cell avidity and tumor recognition: implications and therapeutic strategies. *J Transl Med* (2005) **3**(1):35. doi:10.1186/1479-5876-3-35
26. Rous P, Beard J. The progression to carcinoma of virus-induced rabbit papillomas (shope). *J Exp Med* (1935) **62**(4):523–48. doi:10.1084/jem.62.4.523
27. Baum C, von Kalle C, Staal FJ, Li Z, Fehse B, Schmidt M, et al. Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Mol Ther* (2004) **9**(1):5–13. doi:10.1016/j.ymthe.2003.10.013
28. Fehse B, Roeder I. Insertional mutagenesis and clonal dominance: biological and statistical considerations. *Gene Ther* (2007) **15**(2):143–53. doi:10.1038/sj.gt.3303052
29. Yin Y, Mariuzza RA. The multiple mechanisms of T cell receptor cross-reactivity. *Immunity* (2009) **31**(6):849–51. doi:10.1016/j.immuni.2009.12.002
30. Wu S, Jin L, Vence L, Radvanyi LG. Development and application of phosphoflow as a tool for immunomonitoring. *Expert Rev Vaccines* (2010) **9**(6):631–43. doi:10.1586/erv.10.59

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 July 2013; accepted: 03 November 2013; published online: 21 November 2013.
Citation: Gerdes S, Newrzela S, Glauche I, von Laer D, Hansmann M-L and Roeder I (2013) Mathematical modeling of oncogenesis control in mature T-cell populations. Front. Immunol. 4:380. doi: 10.3389/fimmu.2013.00380

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Gerdes, Newrzela, Glauche, von Laer, Hansmann and Roeder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 03 September 2013
doi: 10.3389/fimmu.2013.00252

Inferring HIV escape rates from multi-locus genotype data

***Taylor A. Kessinger*[1], *Alan S. Perelson*[2] *and Richard A. Neher*[1]*

[1] *Evolutionary Dynamics and Biophysics, Max Planck Institute for Developmental Biology, Tübingen, Germany*
[2] *Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA*

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Viktor Müller, Hungarian Academy of Sciences and Eötvös Loránd University, Hungary
Vitaly V. Ganusov, University of Tennessee, USA
Aridaman Pandit, Utrecht University, Netherlands

***Correspondence:**
Richard A. Neher, Evolutionary Dynamics and Biophysics, Max Planck Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany
e-mail: richard.neher@tuebingen.mpg.de

Cytotoxic T-lymphocytes (CTLs) recognize viral protein fragments displayed by major histocompatibility complex molecules on the surface of virally infected cells and generate an anti-viral response that can kill the infected cells. Virus variants whose protein fragments are not efficiently presented on infected cells or whose fragments are presented but not recognized by CTLs therefore have a competitive advantage and spread rapidly through the population. We present a method that allows a more robust estimation of these escape rates from serially sampled sequence data. The proposed method accounts for competition between multiple escapes by explicitly modeling the accumulation of escape mutations and the stochastic effects of rare multiple mutants. Applying our method to serially sampled HIV sequence data, we estimate rates of HIV escape that are substantially larger than those previously reported. The method can be extended to complex escapes that require compensatory mutations. We expect our method to be applicable in other contexts such as cancer evolution where time series data is also available.

Keywords: HIV, CTL escape, cytotoxic T-lymphocytes, HIV evolution, viral dynamics, selection coefficient

INTRODUCTION

During the first few months of HIV infection, the HIV genome typically undergoes a series of rapid amino acid substitutions that reduce immune pressure by cytotoxic T-lymphocytes (CTLs); this process is referred to as CTL escape (1). The substitutions arise by random mutation and spread through the viral population by impairing either the presentation of viral epitopes on the cell surface or the recognition of the viral epitope by T-cell receptors. Avoiding recognition is an obvious benefit to the mutant virus, but escape mutations can interfere with processes necessary for virus replication and infection and thereby reduce the virus' intrinsic fitness (2–5). The rate at which escape variants displace the founder sequences depends on both "avoided killing" and the fitness cost. To quantify the role of individual CTL clones in controlling the viral population and the fitness costs associated with escape mutations, one would like to infer the escape rate associated with the individual mutations from serially sampled sequence data (4, 6).

With a single escape mutation and dense, deeply sampled data, the escape rate can simply be estimated by fitting a logistic curve to the time course of the mutation's frequency (4, 6). The logistic curve has two parameters: the growth or escape rate and the frequency at the initial time point. In many cases, however, the data obtained from infected patients are scarce, and estimating two parameters reliably from the data is not possible since one needs at least two time points at which the mutation is at intermediate frequency between 0 and 1 (4). **Figure 1** shows an example of such time series sequence data from CTL escape during early HIV infection. Time points are far apart and the sampling depth is low. Furthermore, it is not the case that only a single escape mutation is observed; rather, several mutations rapidly emerge in different places in the viral genome (7, 8). Multiple escapes imply immune pressure on many epitopes. Since the viral population and its mutation rate are large (9, 10), these different escape mutations will arise almost simultaneously. Initially, these escape mutations exist in the population as single mutant genomes until they are combined into multiple mutants by recurrent mutation or recombination (11, 12). The competition between viral variants affects the trajectories of individual escape mutations, so estimating their intrinsic growth rate by logistic fitting is not accurate. This competition is known as "clonal interference" in population genetics. The degree of competition between genotypes depends on the population size, the mutation rate, and the recombination rate in HIV populations. The latter-most is rather low (13, 14), and two strongly selected mutations in a large population are more likely to be combined by additional *de novo* mutation than recombination with another rare single mutation.

Here, we develop a strategy for inference that allows one to obtain robust escape rate estimates from the scarce data typical of studies of CTL escape. The inference is based on explicit modeling of the process of mutation accumulation in the founder sequence. Thereby, we exploit constraints imposed by the underlying dynamics of mutation and selection in the high dimensional space of possible genotypes.

Despite the large number of possible genomes that can be formed from different combinations of escape mutations, we typically observe one or two dominant genotypes at a time – at least during the first few months of the infection. Furthermore, these genotypes dominate only transiently and are quickly displaced by genotypes with an even greater number of escape mutations; see **Figure 1**. These observations agree with results from ref. (15), where a model of acute HIV infection was used to show

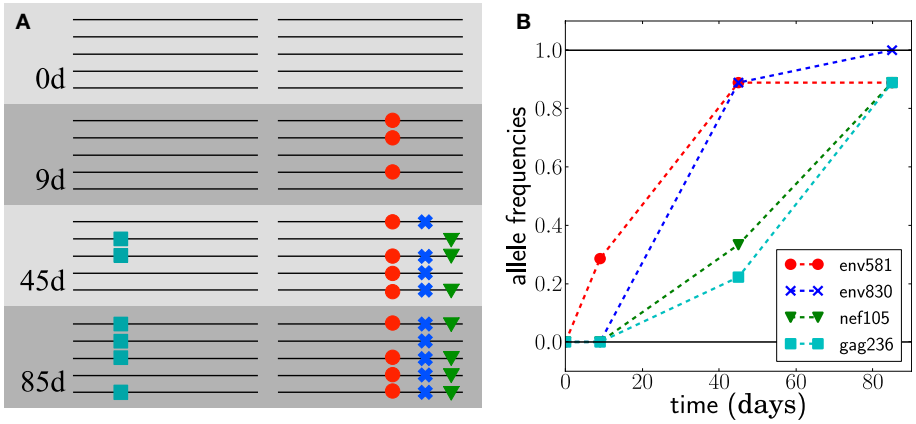


FIGURE 1 | Escape from T-cell mediated immunity. The virus population in patient CH58 quickly acquires four substitutions. **(A)** shows a sketch of genotypes at the first 4 escape mutations, observed at different times; see (7, 8) for the actual data. **(B)** shows the frequencies of the mutations in samples of size 7 at day 9 and size 9 at days 45 and 85.

that strongly selected escape mutations fix sequentially. Note that we don't assume a particular sequence of dominant genotypes *a priori*. Instead, we observe a sequence of dominant genotypes and try to infer the evolutionary scenario that most likely gave rise to this sequence of genotypes. While we model only these genotypes, many minor variants certainly exist. But only those dominant variants that are likely to give rise to the future populations need to be modeled accurately. Later in infection, the viral population is very diverse and cannot be analyzed using our method.

Given a data set from early infection, it is typically straightforward to define a series of dominant genotypes that likely have arisen through step-wise accumulation of mutations. Note that most likely all escape mutations constantly arise in different combinations, but typically only one combination rises quickly enough to dominate the population. This dominant genotype is then in most cases the source for the next dominant genotype. Later in infection, however, recombination is sufficiently frequent that no dominant genotype exists and mutations can spread simultaneously.

In Ganusov et al. (11), a framework for multi-locus modeling of CTL escape is presented. Building on this framework, we explicitly model the transition from one dominant genotype to another, which is a good approximation of the dynamics for rapid CTL escape in acute infection. The restriction to dominant genotypes captures the interference between escapes at different epitopes while avoiding the need to solve the full multi-locus problem.

We will first define a model of the dynamics of escape mutations. This model serves a twofold purpose: it defines the parameters we would like to estimate from the data and provides us with a computational tool to investigate how the accuracy of the inference depends on sampling depth and frequency, as well as how sensitively it depends on the values of parameters such as mutation rates or the population size. We reanalyze existing CTL escape data and find that accounting for multi-locus effects in a finite population results in higher estimates of the escape rates.

RESULTS

MODEL

In the majority of sexually transmitted HIV infections, a single "transmitted/founder" virus initiates the new infection resulting in an initially homogeneous viral population (8, 16). However, as HIV replicates in its new host, mutations accumulate. Mutations within or in proximity to CTL epitopes can reduce immune pressure by facilitating the avoidance of CTL recognition. While one often observes several escape mutations within a single epitope (17, 18), we do not differentiate between different mutations within the same epitope and model L epitopes that can be either be mutant or wild-type. Assuming that the escape at multiple epitopes has additive effects, ε_j, the growth rate (birth rate minus death rate) of a genotype is given by

$$F(g,t) = F_0(t) + \sum_i \varepsilon_i s_i \qquad (1)$$

where $g = \{s_1, \ldots, s_L\}$ specifies the genotype. Here, $s_i = 0$ corresponds to a wild-type epitope at locus i, whereas $s_i = 1$ signifies escape at that epitope. $F_0(t)$ accounts for a genotype independent modulation of the growth rate. The latter could, for example, be due to variable numbers of target cells (19, 20). $F_0(t)$ controls the total population size, while the differences between genotypes are accounted for by $\sum_i \varepsilon_i s_i$ and result in differential amplification of some genotypes over others. The ε_i are the escape rates that we would like to estimate from the data and should be interpreted as the net effect of avoided killing and the possible fitness costs associated with the mutation; see e.g., Ganusov et al. (11). The fitness costs are modulated by the overall growth rate of the viral population and could therefore be slightly time dependent. We neglect this complication. Within our model, mutations arise at a rate μ per base per generation. This rate can be epitope dependent. Motivated by the frequent template switching of HIV reverse transcriptase (21), our general model of the HIV population includes recombination, which is assumed to occur with rate r. In the event

of recombination, all L epitopes are reassorted, but an explicit genetic map could be implemented as well.

We implemented our model as a computer simulation in Python using the population genetic library FFPopSim (22). The simulation stores the population $n(g, t)$ of each of the 2^L possible genotypes. In each generation, the expected changes of the $n(g, t)$ due to mutation, selection, and recombination are calculated. The population of the next generation is then sampled from the expected genotype frequencies $\gamma(g, t) = n(g, t)/N$. The size of the population, N, can be set at will each generation. In this way, up to 15 epitopes can be simulated for 1000 generations within seconds to minutes.

A typical realization of the population dynamics is shown in **Figure 2**, where we have assumed a generation time of 1 day. As expected, the population is dominated by one genotype at a time. Furthermore, the mutations accumulate in decreasing order of escape rate, and the new dominant genotype arises from the previous by incorporation of the mutation with the largest escape rate available. There are, however, many minority genotypes which are rarely observed. **Figure 2C** shows the frequencies on a logarithmic scale, where the minor variants are visible. We use these simulations to test the accuracy and robustness of the inference procedure developed below.

Of the many possible genotypes that are present at any moment, only a small fraction is likely to be observed in a small sample and to be relevant in the future. Simulations and data suggest that the dominant genotypes accumulate mutations one by one – this greatly simplifies the task of estimating escape rates from the data. Instead of considering the dynamics of all possible genotypes (2^L), we will restrict the inference to a chain of genotypes, each containing one additional mutation compared to its predecessor.

The best estimate for the HIV generation time is $d = 2$ days (23), while estimates of escape rates are typically given in units of inverse days rather than generations. For simplicity, we simulate our model assuming one generation per day and state all rates in units of 1/day. Our results are insensitive to the choice of the generation time. Doubling the generation time has similar effects to dividing the population size by 2, as this keeps the strength of genetic drift constant.

INFERRING THE ESCAPE RATES

Suppose we have obtained sequence samples of size n_i at different time points t_i and each of these samples consists of different genotypes g present in $k(g, t_i)$ copies. If the actual frequencies of those genotypes at different times are $\gamma(g, t_i)$, the probability of obtaining the sample at t_i is given by the multinomial distribution

$$P\ (\text{sample}) = \frac{n_i!}{\prod_g k\,(g, t_i)!} \prod_g \gamma\big(g, t_i\big)^{k(g,t_i)} \qquad (2)$$

If the underlying dynamics was deterministic, the frequencies $\gamma(g, t)$ would be unique functions of the model parameters we want to estimate. In that case we could use Bayes' theorem, choose suitable priors, and determine the posterior distribution of the parameter values. However, both the model and the actual viral dynamics are stochastic, and "replaying" the history would result in different trajectories. Furthermore, most of the 2^L possible genotypes remain unobserved. This leaves us with the choice of either some type of approximate Bayesian computation that compares repeated simulations of the model with appropriate summary statistics (24) or a reduced description of only the observed genotypes, with the stochasticity captured by nuisance parameters (25).

We opt for the latter and model only those genotypes that dominate the population. We label these genotypes by the number of escape mutations they carry, e.g., g_1 carries the first escape mutations, g_2 the first and the second, and so forth. The frequency of a genotype is affected by stochastic forces only while it is very rare. If the genotype is favored, it will rapidly rise to high frequency, and the stochastic effects will no longer be relevant. It is therefore convenient to summarize the stochastic behavior by the time, τ, at which its frequency crosses the threshold to essentially deterministic dynamics. Since the dynamics is deterministic after this "seed

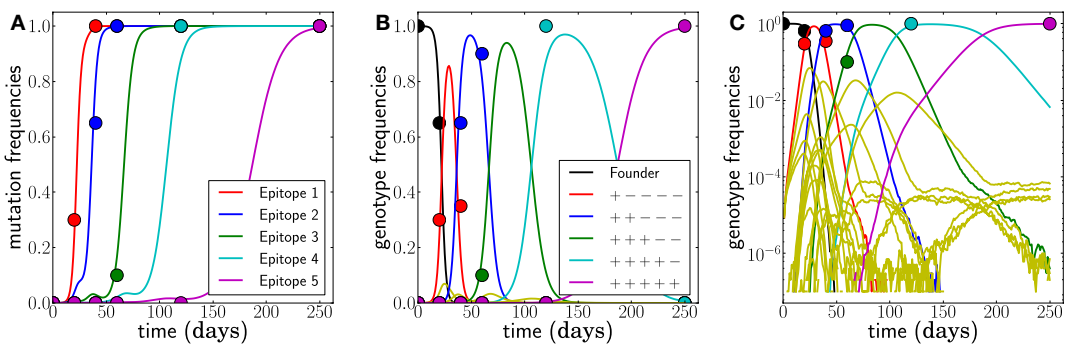


FIGURE 2 | Example of simulated escape mutations spreading through the population. **(A)** Even though all epitopes are targeted from $t = 0$, escape mutations spread sequentially. The mutation frequency in a sample of size 20 at different time points is indicated by colored dots. **(B)** The rising mutation frequencies are associated with the rise and fall of multi-locus genotypes. The founder virus is first replaced by a dominant single mutant, which itself is replaced by a double mutant and so forth. Note, however, that the virus population explores many combinations of mutations but that these minor variants never reach appreciable frequency. This is best seen in **(C)**, where all 32 genotype frequencies are shown on a logarithmic scale. These rare variants are rarely sampled, and their noisy dynamics suggests that little information can be gained from them. Here, $N = 10^7$, $\mu = 10^{-5}$, and $r = 0$, and escape rates are $\varepsilon_i = 0.5$, 0.4, 0.25, 0.15, 0.08 per day.

time," all the (unobserved) stochasticity can be accounted for by an appropriate choice of the seed time (26, 27). For each of the dominant escape variants, g_j, with $j = 1$ to $j = L$ escaped epitopes, we define a seed time τ_j to accommodate the stochastic aspects of the escape dynamics.

After crossing the deterministic threshold, the population frequencies of the dominant genotypes evolve according to

$$\dot{\gamma}_j(t) = F(g_j, t)\,\gamma_j(t) + \mu\left[\gamma_{j-1}(t) - \gamma_j(t)\right] \qquad (3)$$

if $t > \tau_j$. Conversely, $\gamma_j(t) = 0$ for $t < \tau_j$. The growth rate $F(g_j, t)$ of genotype j is the sum of the escape rates ε_k of the epitopes $k = 1, \ldots, j$ and the density regulating part $F_0(t)$; compare to equation (1). The escape rates are what we would like to estimate. The seed time, τ_j, corresponds to the time at which a genotype with all escape mutations up to mutation j first establishes[1]. At the seed time, we initialize the genotype frequency at $\gamma_j(\tau_j) = N^{-1}$. If seed times are chosen appropriately, this model provides a very accurate description of the frequency dynamics of the dominant genotypes in the full stochastic model; see **Figure 3**.

At face value, the deterministic model has two parameters per epitope – one escape rate and one seed time. The seed times, however, are quite strongly constrained by basic facts of the evolutionary dynamics. The genotype g_j carrying mutations $i = 1, \ldots, j$ arises with rate $\mu N(t)\gamma_{j-1}(t)$ from the genotype g_{j-1} carrying only $j-1$ mutations. This means it is unlikely that genotype j arises early while $\gamma_{j-1}(t)$ is still very small. However, once the previous genotype $j-1$ is common, genotype j is produced frequently. The distribution of the time at which the first copy of genotype j arises is given by the product of the rate of production and the probability that it has not yet been produced. The latter is the negative exponential of the integral of the production rate up to this point. Hence, the distribution of the seed time τ_j, given the trajectory of the previous genotype γ_{j-1}, is given by

$$Q(\tau_j|\gamma_{j-1}(t)) \approx \mu N(\tau_j)\gamma_{j-1}(\tau_j) e^{-\mu \int_0^{\tau_j} N(t)\gamma_{j-1}(t)\,dt}. \qquad (4)$$

Since the $\gamma_j(t)$ are uniquely specified by $\{\tau_k, \varepsilon_k\}_{k=1,\ldots,L}$, we can write the posterior probability of the parameters as

$$P\left(\{\varepsilon_j, \tau_j\}\right) \propto \prod_i P\left(\text{sample}_i|\Theta\right) \prod_j Q\left(\tau_j|\Theta\right) U\left(\varepsilon_j\right), \qquad (5)$$

where $\Theta = \{\varepsilon_k, \tau_k\}_{k=1\ldots L}$ and $U(\varepsilon_j)$ is our prior on the escape rates. We employ a Laplace prior $U(\varepsilon) = \exp(-\Phi\varepsilon)$ parameterized by Φ favoring small escape rates. The prior regularizes the search for the minimum and results in conservative estimates of escape rates.

OBTAINING MAXIMUM LIKELIHOOD ESTIMATES

Finding the set of escape rates and seed times that maximizes the posterior probability can be difficult due to multiple maxima and ridges in the high dimensional search space, and uncertainty remains. To ensure that the global optimum will be reliably discovered, we exploit the sequential nature of the dynamics and use the fact that earlier escapes strongly affect the timing of the later ones, but not vice versa. Thus adding genotypes with an increasing number of mutations one at a time results in a reasonable initial guess on top of which a global true multi-locus search can be performed.

We have implemented such a search in Python, and the computationally expensive calculation of the posterior probability is implemented in C. The code infers parameters as follows:

- Fit the first escape assuming $\tau_1 = 0$ by a simple one dimensional minimization. This assumes that single mutants are already present in the population, consistent with the large viral population size present by the time a patient has been identified as HIV-1 infected (28, 29).
- Add additional epitopes successively by mapping the entire two-dimensional posterior distribution $P(\varepsilon_j, \tau_j)$ at fixed $\{\varepsilon_k, \tau_k\}$ for $k < j$. This step is illustrated in **Figure 4A**.
- Refine the estimates through local optimization via gradient descent, Monte Carlo methods, or local exhaustive search. The resulting parameters and trajectories are shown for one example in **Figure 4B**.
- Generate posterior distributions by Markov chain Monte Carlo (MCMC).

This procedure is described in more detail in the Section "Materials and Methods." Fitting five epitopes takes on the order of a minute on one 2011 desktop machine (Apple iMac i7 2.93 GHz). Generating the local posterior distribution by MCMC takes roughly 20 min for 10^6 steps.

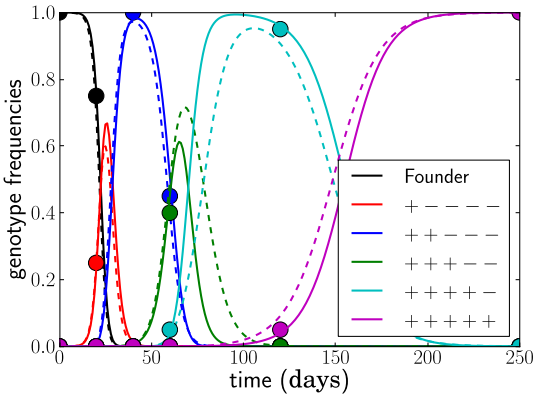


FIGURE 3 | The deterministic model parameterized by seed times τ_j for the L dominant genotypes and the escape rates of epitopes ε_i (solid lines) captures the dynamics of the stochastic model accurately (dashed lines). The trajectories (and seed times) vary from run to run. In this run, $N = 10^7$, $\mu = 10^{-5}$, and $r = 0$ and the escape rates are $\varepsilon_j = 0.5$, 0.4, 0.25, 0.15, 0.08 per day.

COMPARISON TO SIMULATED DATA

To evaluate the accuracy and reliability of our inference scheme, we performed true multi-locus stochastic simulations using

[1]There is a brief period after the initial production of the mutation during which the dynamics is stochastic and the initial mutant establishes only with a probability roughly equal to εd, where $d = 2$ days is the generation time (23). However, we find $\varepsilon d \approx 1$ and ignore this complication.

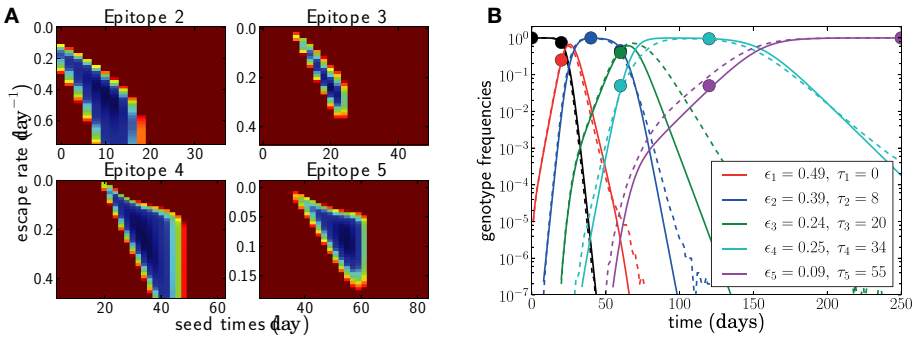


FIGURE 4 | Adding epitopes one by one is a feasible and reliable fitting strategy. Assuming we know the population was homogeneous at $t = 0$, there is only one free parameter for the first epitope, which is easily determined. For all subsequent epitopes, we need to determine the seed time τ_j and the escape rate ε_j. In **(A)**, the negative log posterior probability of these parameters is shown for each of the epitopes. The surface typically exhibits a single minimum. **(B)** shows the genotype frequencies of the founder virus and the dominant escape variants (solid lines: model fit, dashed lines: actual simulated trajectories). The estimated escape rates of individual epitopes and the seed times of genotypes containing all escape mutations up to j are given in the legend. Only the samples indicated by balls (20 sequences at each time point) were used for the estimation. In this run, $N = 10^7$, $\mu = 10^{-5}$, $r = 0$, and the escape rates $\varepsilon_j = 0.5, 0.4, 0.25, 0.15, 0.08$ per day.

FFPopSim (see Materials and Methods) and sampled genotypes from the simulation at a small number of time points. Time points and sample sizes were chosen to mimic patient data. We then inferred parameters from this "toy" data set and compared the result to the actual values. When interpreting these comparisons, it is important to distinguish two sources of error. First, limited sample size and sampling frequency will incur errors due to inaccurate estimates of the actual genotype frequencies from the sample. The second source of uncertainty is an inappropriate choice of model or model parameters. Such inappropriate model choices might include wrong estimates of the population size or mutation rates, the presence or absence of recombination, or time variable CTL activity.

We generate data assuming escape rates $\varepsilon_j = 0.5, 0.4, 0.25, 0.15, 0.08$ per day and sample the population on days $t_i = 0$, 20, 40, 60, 120, 250. An example of such samples is shown in **Figure 2**. Note that each genotype is typically only sampled at a single data point; it easily happens that a genotype is hardly seen at all. We therefore expect all inferences to be quite noisy as is the case with patient data.

Sample size and sampling frequency dependence

With more frequent and deeper sampling, inferring the model parameters is expected to become simpler. Indeed, as soon as each genotype is sampled more than once at intermediate frequency, one can estimate its growth advantage simply from its rate of increase. This is the rationale behind previous studies such as (4, 6). In many data sets, however, this condition is not met. By constraining the seed time based on the evolutionary trajectory of the previous escape, our method is able to produce a more accurate reconstruction of parameters with less data.

Figure 5 shows the estimates obtained as a function of the sampling frequency and sample size. Increasing the sample size improves the estimates only moderately, whereas increasing the sampling frequency leads to substantial improvements.

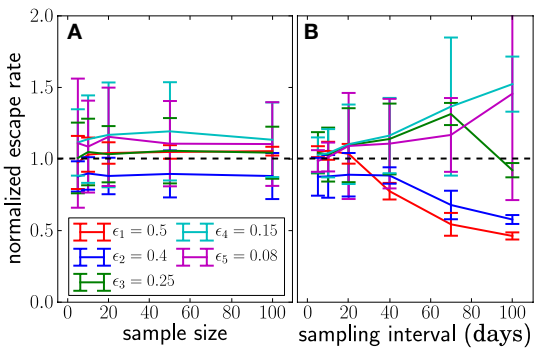


FIGURE 5 | The dependence of the accuracy of inference on sample sizes (A) and sampling intervals (B). The actual normalized escape rate is 1.0 and is shown by the dashed line. Sample size only moderately affects the accuracy, while sparse sampling (every 40 days in this example) leads to serious loss of accuracy. Sample size is $n = 20$ when sample intervals are varied, and sampling times are as illustrated in **Figure 2** when sample size is varied. The plots show the mean ± one standard deviation. The actual values of the escape rates simulated are shown in the legend (same on both panels). In each run, $N = 10^7$, $\mu = 10^{-5}$, and $r = 0$. Mean and standard deviation at each point are calculated from 100 independent simulations.

Model deviations

The population size and the mutation rate explicitly enter our model through the seed time prior, but we rarely know these numbers accurately. Hence we need to understand how inaccurate assumptions affect our estimates. If we assume that $N\mu$ is larger than it really is, our inference method will favor seeding subsequent genotypes too early, which in turn results in erroneously small estimates of escape rates. We varied N and μ and observed the expected effect on the estimates as shown in **Figure 6**. The dependence on μ is stronger than that on N, since the effect of a larger population size is partly canceled by the longer time necessary to amplify the novel mutation to macroscopic numbers.

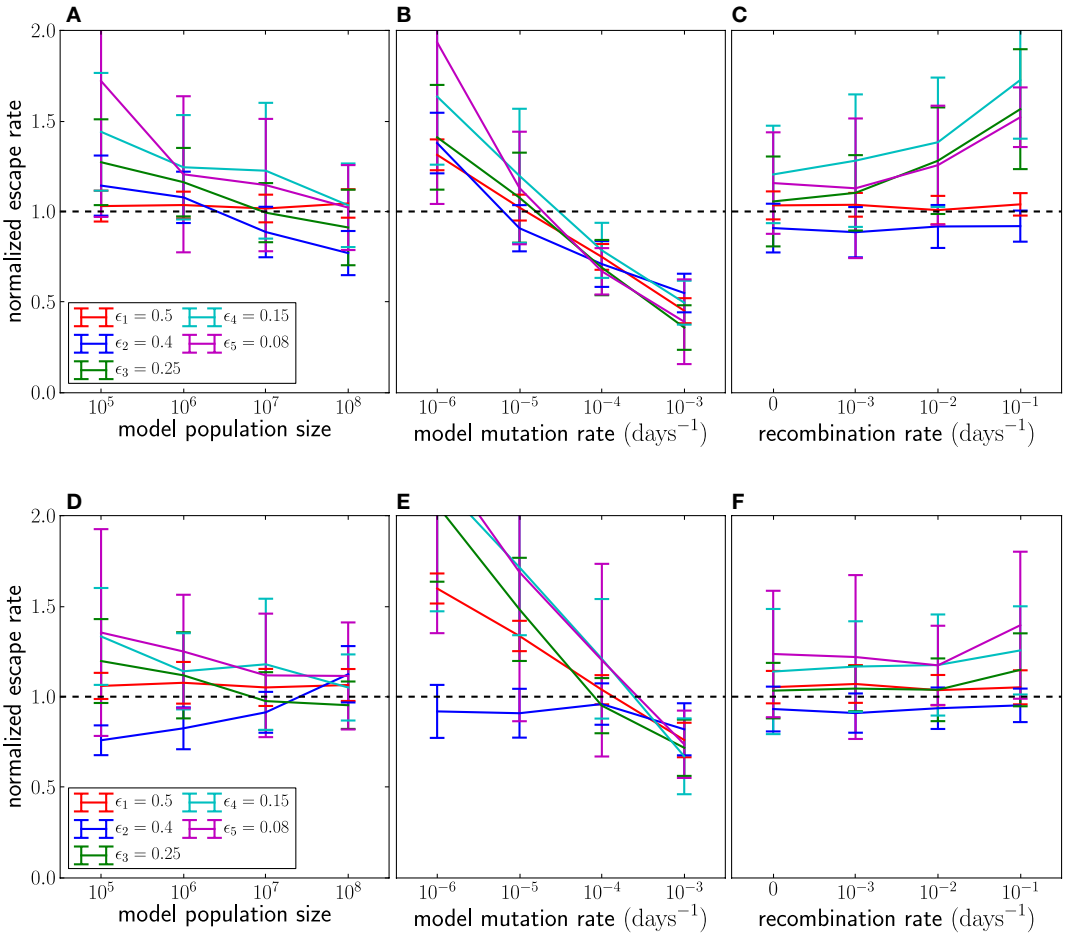


FIGURE 6 | The effect of assuming the wrong population parameters on the escape rate estimates. To quantify the robustness against wrong assumptions, we simulate escape dynamics with parameters different from those assumed in the escape rate estimation. **(A–C)** show simulations with $N = 10^7$ and $\mu = 10^{-5}$ per day, while **(D–F)** use a 10-fold higher mutation rate $\mu = 10^{-4}$. In **(C,F)**, the simulated recombination rate varies as shown. **(A,D)** Assuming a too small population size results in estimates that are too large. The effect is more pronounced at lower mutations rates. **(B)** Similarly, if the mutation rate is assumed too large, the estimated seeding of multiple mutants occurs too early and the estimates of escape rates are too low. Note that assuming the correct rates [$\mu = 10^{-5}$ in **(B)** and $\mu = 10^{-4}$ in **(F)**] results in unbiased estimates. **(C,F)** If the population recombines, the actual seed times are smaller than those estimated by the fitting routine. To compensate for the shorter time interval during which the escape variant rises, the estimates of escape rates are larger than the actual escape rates, at least at low mutation rates. For high mutation rates, recombination is less important because additional mutations are more efficient at producing multiple mutants than recombination. Mean and standard deviation at each point are calculated from 100 independent simulations.

However, even the dependence on μ is rather weak, and changing μ 10-fold only changes estimates of escape rates by $\pm 50\%$. The underlying reason is that the seed times depend primarily on the logarithm of $N\mu \cdot Q(\tau_j | \gamma_{j-1}(t))$ (see equation (4)), which peaks when $N\mu\gamma_{j-1}(t) \approx 1$. Because $\gamma_{j-1}(t)$ is growing exponentially, the position of the peak changes only logarithmically with the prefactor $N\mu$. Changes in μ also affect the dynamics through the initial rise in frequency of novel genotypes due to recurrent mutations; see equation (3).

Another factor that affects seed times is recombination. HIV recombines via template switching following the coinfection of one target cell by several virus particles (21). In chronic infection, coinfection occurs with a frequency of about 1% (13, 14). Recombination is not modeled in the seed time prior of our inference method but can speed up escape by combining escape mutations at different epitopes. As a result, if recombination is present, seeding tends to happen earlier than our prior would suggest. If the model assumes that seeding occurs later than in reality, there is less time for an escape variant to grow to its observed frequency. Hence the estimated escape rate (growth rate) is larger than the actual escape rate to compensate for the shorter time. In **Figure 6**, we compare the estimates obtained by applying our inference method to simulation data with recombination. Recombination starts to have substantial effects once coinfection exceeds a few percent. Recombination primarily affects the incorporation of more weakly selected mutations and can be ignored for very strongly selected CTL escape mutations. Recombination also has negligible effects if the mutation rate is large as is seen in **Figure 6F**.

Unobserved intermediates and compensatory mutations

The time intervals between successive samples are sometimes too large to observe the accumulation of single mutations, so the dominant genotype at one time point differs by more than one mutation from the previous. This can arise for two reasons. First, one or several unobserved genotypes may have transiently been at high frequency but been out-competed by later genotypes before the next sample was taken. Second, one escape might have required more than one mutation, for example because single mutants are not viable and a compensatory mutation is needed (30). Both scenarios can be accounted for in our scheme and are illustrated in **Figure 7**.

Unobserved, but individually beneficial, intermediate genotypes can be included by assuming they all have the same escape rate and were seeded one from the other. There is not sufficient information to estimate more than an average escape rate for all of them. For a given set of sampled frequencies, the estimated escape rates increase as more and more intermediates are assumed. Such unobserved intermediates are common in the data from infected individuals analyzed below.

Compensatory mutations and "multiple-hit" escapes can be accounted for by replacing the single site mutation rate in equation (4) by the effective rate at which the viable escape mutant appears. In the simplest case where all intermediate states are lethal and mutations are independent, this rate is simply the probability μ^k, where k is the number of mutations needed. In other cases, the rates to multiple hits can be calculated using branching process approximations (31, 32). The choice of the relevant effective mutation rate for complex escapes must be made on a case-by-case basis. The effective mutation rate of a multiple-hit escape will often be low enough that its seed time is not very well constrained. If, for example, the population size is $N = 10^8$ and the effective mutation rate is 10^{-10}, the seed time distribution has a width of more than 100 days. Given this weak constraint, more data are required in order to estimate the escape rate accurately; see **Figure 7**.

IMMUNE ESCAPE IN HIV-INFECTED PATIENTS

Cytotoxic T-lymphocyte escape was characterized in detail in the patients CH58, CH40, and CH77 (7, 8) and further analyzed in Ganusov et al. (4). Sequences were obtained by single genome amplification followed by traditional sequencing. The data are sparser and less densely sampled than most of the artificial examples analyzed above, so any estimates are necessarily rather imprecise. Furthermore, we do not know exactly when infection occurred or CTL selection started. The days given in the above papers are relative to the date of identification of the patient as HIV infected. It has been estimated that in a chronically infected patient, there are a total of around 4×10^7 infected cells (33). Hence, the population size is $N \approx 10^7$ but might be larger during peak viremia or smaller due to bottleneck effects or the myriad of factors influencing patient-to-patient variation in viral load. We determined posterior distributions for population sizes ranging from $N = 10^5$ to $N = 10^8$. The mutation rate was set to 10^{-5} per day (10). This value is appropriate if only one escape mutation per epitope is available. If escape can happen in many different ways, a higher rate of about $\mu = 10^{-4}$ per day should be used, so we repeated the estimation with $\mu = 10^{-4}$ finding similar results; see below. Both of these scenarios are observed (18). Recombination in HIV occurs but is not modeled here because its rate is low (13, 14), and it is expected to be less relevant for the strong escapes in large populations. In large populations, recurrent mutation is often more effective at accumulating escape mutations than recombination between two rare variants. Nevertheless, the neglect of recombination can lead to overestimation of escape rates; see above. Lastly, we assume that infection occurred $\tau = 20$ days before the patient was identified and the viral population sampled (7).

For each patient, we initially considered all non-synonymous mutations that are eventually sampled at high frequency as potential candidates for sequential escape mutants. Nearby mutations in the same epitope were combined into one escape. We refined this list of candidates by considering only time points early in infection

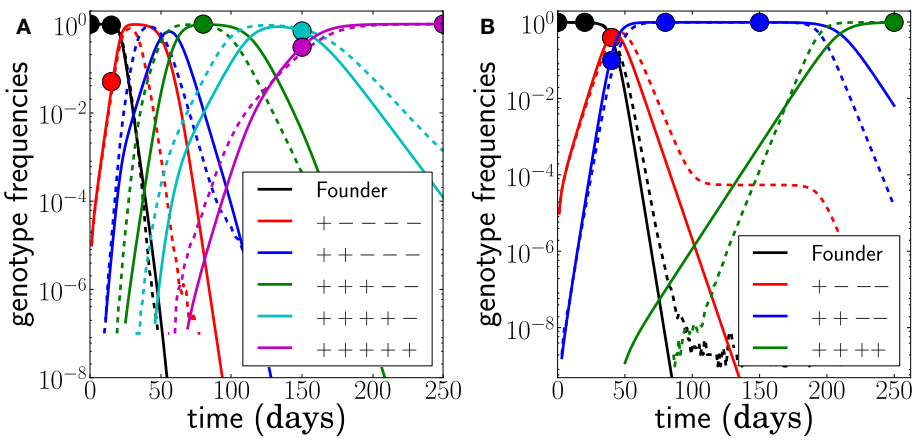


FIGURE 7 | Unobserved intermediates and compensatory mutations. **(A)** shows a scenario where the genotype with only 2 escape mutations (blue) was not observed even though this genotype was transiently at high frequencies. We fit this scenario by assuming both mutations have the same escape rate but occur sequentially ($N = 10^7$, $r = 0$, $\mu = 10^{-5}$). **(B)** shows a scenario where escape mutations 3 and 4 only occur together and any genotype containing only one of the two mutations is not viable. Hence the effective mutation rate into the genotype is $\mu^2 = 10^{-10}$ and the waiting time for this genotype is longer. Note that the population size is $N = 10^9$ in this example ($r = 0$, $\mu = 10^{-5}$). The last escape only appears once the previous escape mutations have reached frequency one, and the seeding time is quite variable.

that were sampled with more than 5 genomes per time point and only the earliest 3-6 strong escapes. All samples used had between 7 and 15 sequences. The frequencies of these escape mutations and their linkage into multi-locus genotypes in the 5′ and 3′ half of the genome, which were sequenced independently, can be easily determined from the alignment provided in Salazar-Gonzales et al. (8). Linkage information between the 5′ and 3′ half genomes is missing but can in all cases be imputed using the assumption of sequential escapes. We ignored mutations whose frequency does not increase monotonically such as pol80 in subject CH40. Later in infection, there is extensive non-synonymous diversity and it is not feasible to fit a time course for most of these mutations.

In CH40 we considered samples at time points $t = 0, 16, 45, 111$, and 181 days and identified escape in six epitopes; the first escape occurs in nef185, followed by three indistinguishable escapes at gag113, gag389, and vpr74 and two additional escapes in vif161 and env145. Following Ganusov et al. (4) the number in the epitope name refers to the beginning of the 18-mer peptide covering the epitope. The mutation at env145 was not analyzed in Ganusov et al. (4), and 145 is simply the number of the mutated amino acid in gp120. The indistinguishable escapes gag113, gag389, and vpr74 are treated as described in the section on unobserved intermediates (all three escapes are assumed to have identical escape rates and only their seed times are varied). Note that the fifth escape at epitope vif161 shows almost the same escape pattern as the three indistinguishable escapes preceding it. The escape rates of gag113, gag389, vpr74, and vif161 should therefore be interpreted with care. In CH58 we considered samples at time points $t = 0$, 9, 45, and 85 days and identified four escapes; the first escape is at env581 and the second at env830, followed by nef105 and gag236. In CH77 we considered samples at time points $t = 0$, 14, and 32 days and identified four escapes, namely the first escape in tat55 and subsequent escapes in env350, nef17, and nef73.

Given the above assumptions, we obtained estimates for the seed time and escape rate of each mutation. For each patient, we obtained initial estimates using a naïve single epitope fit for each mutation; then, we iterated our multi-epitope fitting model five times. Next, we obtained posterior distributions for the escape rates, all shown in **Figure 8**, by performing a MCMC simulation using the likelihood function given in equation (5). After obtaining our estimates, we randomly changed the escape rates in increments of ± 0.01 and the seed times by ± 1, reevaluated the likelihood, and accepted the change with probability $\min(1, \exp(\Delta))$, where Δ is the change in log-likelihood. The resulting Markov chain was run for 10^6 steps with samples taken every 1000 steps.

Figure 8 shows the posterior distributions of the escape rates for different epitopes in the three patients evaluated assuming a mutation rate $\mu = 10^{-5}$ per day. Larger population sizes result in smaller estimates of the escape rates, as expected from **Figure 6A**. The posterior distributions for the first escapes are often very tight, but they depend on the time of the onset of CTL selection, which we have set here to $T = 20$ days prior to the first sample. If we assume that the time of the onset of CTL selection coincided with the first sample (i.e., $T = 0$), the estimates of escape rates of the first epitope ϵ_1 are around 0.9, while later escapes are almost not sensitive to the choice of T.

While the posterior distributions of escape rates of subsequent escape rates are quite broad, they nevertheless suggest that escape rates can be substantially higher than previously estimated (4, 6). Furthermore, the escape rate is not obviously negatively correlated with the time of emergence during acute infection with HIV-1, at least for the earliest four to six escapes. The underlying reason for this is that selection on a late escape is only active after the successful multiple mutant has been produced. In previous single epitope estimates, selection was allowed to act on the mutant frequency from the very beginning, resulting in a reduced estimate of the escape rate. **Figure 8** also shows the inferred trajectories for the most likely parameter combination for patient CH40. One clearly sees the rapid rise and fall of multiple genotypes between the second and third time point. Given the large number of genotypes involved and the little data available, the escape rates estimated for this case are rather noisy. But this analysis clearly shows that strong selection is necessary to bring four mutations to fixation in just a few weeks. We repeated the analysis of the patient data assuming a mutation rate of $\mu = 10^{-4}$ and show the results in **Figure 9**. The overall picture is similar to what we found for $\mu = 10^{-5}$ per day, but escape rates tend to be lower.

DISCUSSION

We have suggested a way to infer viral escape rates from time series data sparsely sampled from the evolutionary dynamics of asexual or rarely sexual populations such as HIV. We exploit the sequential manner in which escape mutations accumulate, which allows us to constrain the times at which new escape mutations arose. These constraints regularize the inference to a large extent, but additional stability is gained by prioritizing small escape rates through an exponential prior.

Escape rates of single escape mutations have so far been estimated by comparing the time series data to a model that assumes logistic growth of the mutation with a constant rate. This approach has been used to analyze the intra-patient dynamics of recombinant HIV (34), drug resistance (35, 36), and CTL escape dynamics (4, 6, 20, 37, 38). While these methods work well if each mutation is sampled multiple times at intermediate frequencies, they provide very conservative lower bounds when data are sparse. Furthermore, they ignore the effects of competition between escapes at different epitopes and assume that each epitope can be treated independently. Since the recombination frequency in HIV is low (13, 14, 39), this can be a poor approximation. Our method improves on previous methods on both of these counts. We explicitly model the competition between escape mutations. This competition places constraints on the times at which genotypes with multiple escapes first arise (double mutants arise only after the single mutants), which makes the inference more robust and the lower bound tighter.

A related method to estimate CTL escape rates has been proposed by Leviyang (12), who modeled multiple escape mutations by an escape graph that is traversed by the viral population. Combining these two approaches, intra-epitope competition as modeled in (12) and the between epitope competition studied here, would be an interesting extension. Similar ideas have been developed in the context of mutations in cancer or evolution experiments (40).

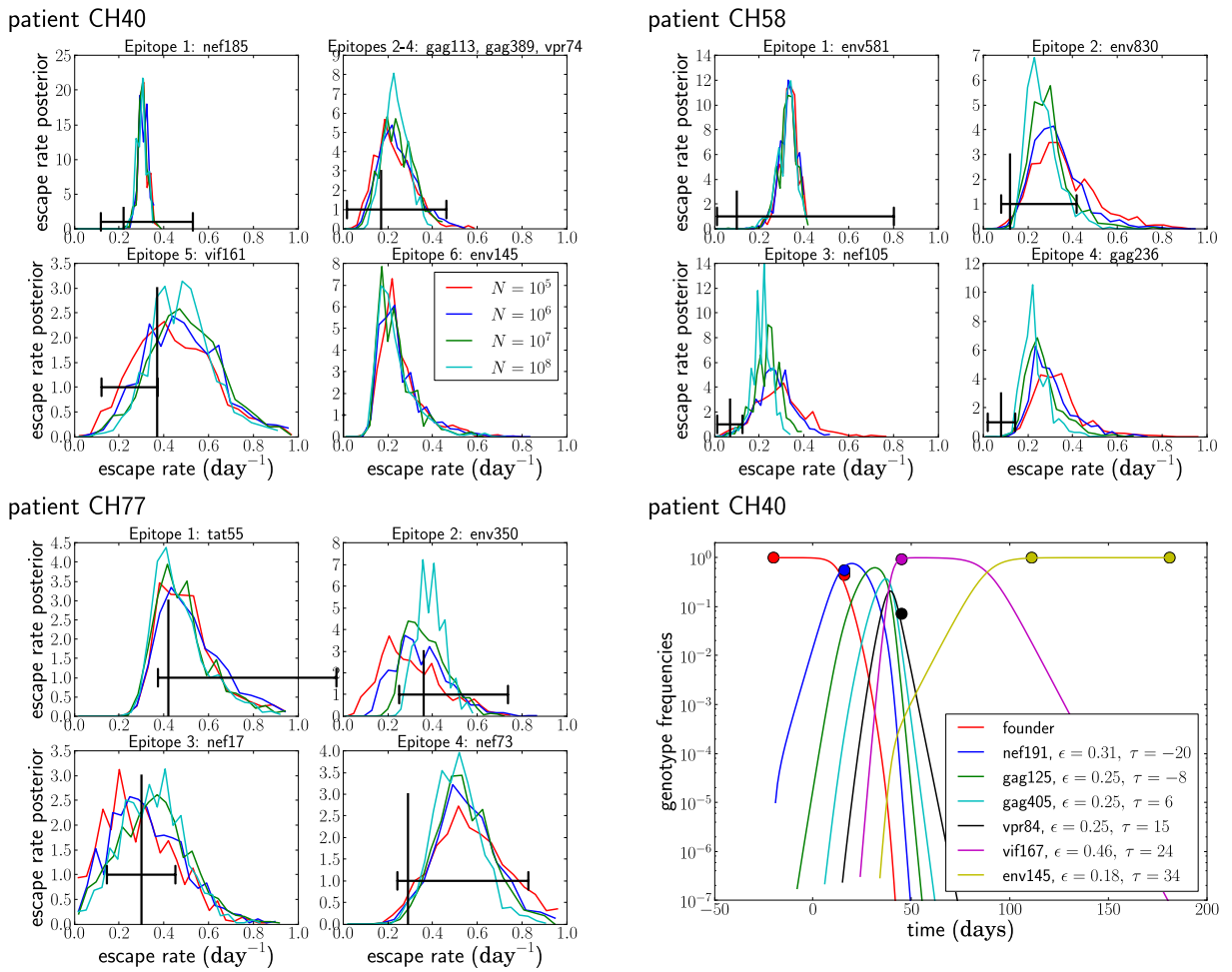


FIGURE 8 | The posterior distribution of the escape rates for different population sizes. It is assumed that CTL selection starts 20 days prior to the date of identification, and the fitness prior has weight $\Phi = 10$. The black vertical and horizontal lines indicate the estimates and confidence intervals obtained in Ganusov et al. (4). Note that the mutation env145 in CH40 was not analyzed in Ganusov et al. (4). The lower right panel shows the most likely genotype trajectories for patient CH40 with parameters $N = 10^7$ and $\mu = 10^{-5}$. Each curve is labeled by an epitope but should be understood as the frequency of the genotype that has escaped at this and all previous epitopes. Note that no data are available to differentiate epitopes gag113, gag389, and vpr74. For those, we assume an arbitrary order and equal escape rates as explained in the section on unobserved intermediates.

While previous methods neglect interactions between epitopes altogether – equivalent to assuming very rapid recombination – our method ignores recombination during the inference. By comparison with simulations that include recombination, we have shown that neglecting recombination can result in overestimation of the escape rates by roughly 30% at plausible recombination rates of 1% (13, 14). We also show that neglecting recombination is less of a problem at higher mutation rates. Note that neglecting recombination cannot explain the larger escape estimates compared to previous studies. For patient CH58 we find escape rates that are up to threefold higher than earlier estimates (4), while we never see such a large deviation in our sensitivity analysis. Furthermore, the errors made when neglecting recombination for rapid early escapes are comparable to the uncertainties that result from infrequent sampling or more severe deviations of the model from reality, such as time variable CTL activity.

Reanalysis of CTL escape data from HIV using our method suggests that CTL escapes are substantially more rapid than previously thought. Even with a large prior against high escape rates ($\Phi = 10$), we estimate that the escape rates of the first 4–6 escapes are on the order of 0.3-0.4 per day. The estimates at large population sizes are fairly insensitive to the prior for population sizes of 10^6 or larger. Early in infection, it is plausible to assume that the relevant size is $N = 10^7$ (28, 29, 41). If population sizes are small, relaxing the prior against high escape rates results in larger estimates, which further supports our finding that escape rates are often large and competition between escapes needs to be modeled. Given the sparse data, we can only estimate parameters of simple models and have to neglect many complicating features of HIV biology. Among other factors, the rate at which escape mutations are selected depends on the overall R_0 of the infection, and CTL selection is probably time variable (4). The

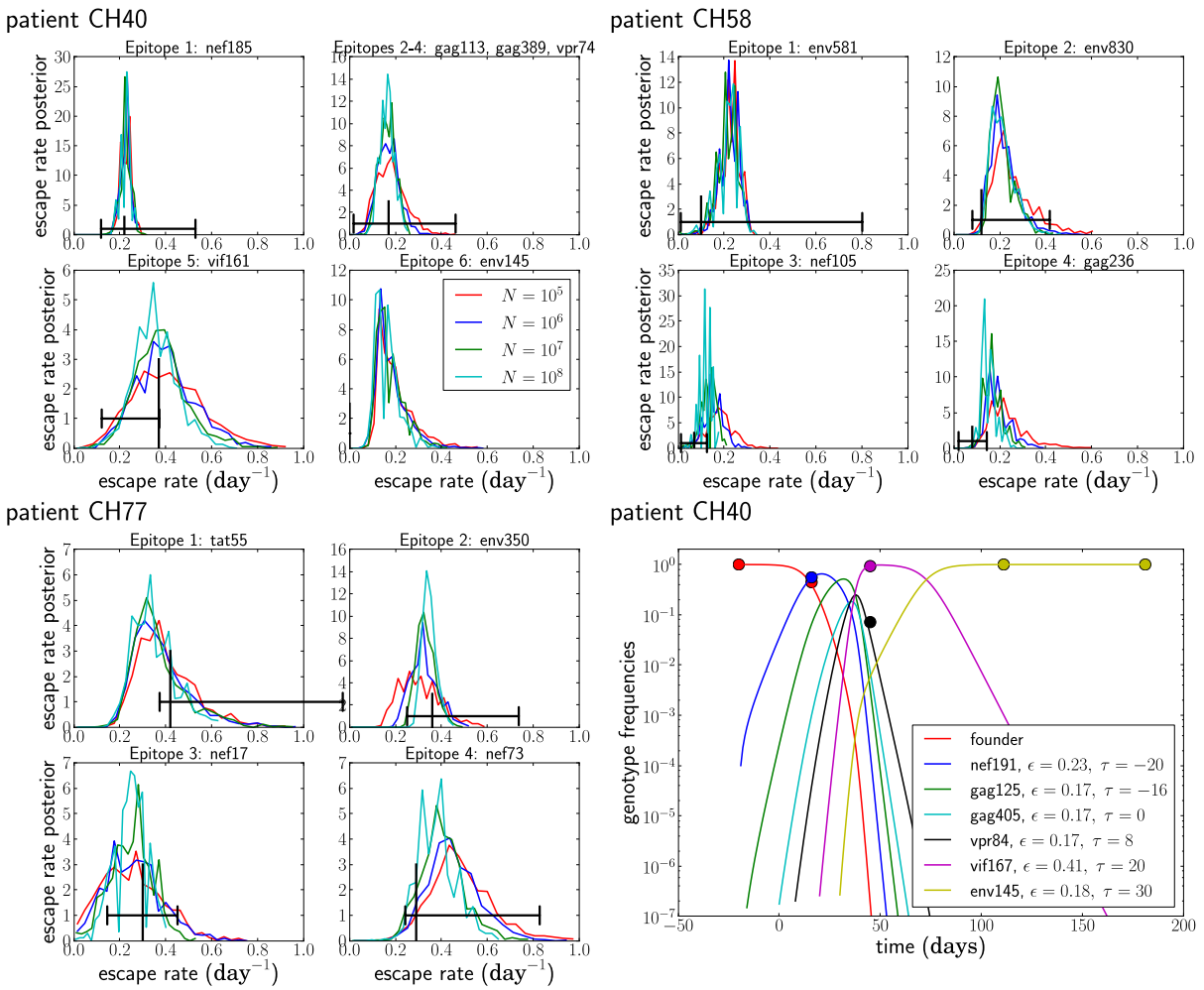


FIGURE 9 | Posterior distributions of escape rate assuming a mutation rate $\mu = 10^{-4}$ per day. See **Figure 8** for other details.

estimated parameters therefore represent time averaged effective escape rates.

The timing of escape has been shown to depend on epitope entropy and immunodominance (42). However, we modeled only the first four to six escapes in each patient, from which rather little information about differential timing can be obtained. In the case of CH77, the first four escapes occurred within a month from the identification of the patient. In patient CH58, it took roughly 3 months for four escapes to spread and the estimated escape rates are lower as expected. In the case of CH40, four of the six escapes show almost or completely indistinguishable escape patterns and we have little power to differentiate the escape rates at epitopes gag113, gag389, vpr74, and vif161. Hence any meaningful correlation with immunological features and epitope sequence conservation, i.e., low entropy, requires more data.

The proposed method to analyze multi-locus time series of adaptive evolution could be useful in many context where the genotypic compositions of large populations of viruses or cells can be monitored over time. Whenever mutations occur rapidly enough that they compete which each other, this competition has to be accounted for in the analysis. Outside of virus evolution, possible applications include the development of cancer and microbial evolution experiments.

MATERIALS AND METHODS

DATA PREPARATION

Our fitting method uses counts k_{ij} of genotypes g_j at time points t_i to infer escape rates of individual mutations. The procedure used to obtain successive genotype counts from sequence data sampled from patients is outlined in the text. As input data, our analysis scripts expect a white-space delimited text file with a format shown in **Table 1**. In addition, a separate file with the total number of sequences at each time point can be provided. This file is expected to have the same format as the matrix with the genotype counts; see **Table 1**. In absence of such a file, the sample sizes at each time point are obtained by summing the genotype counts.

To test our method, artificial data $k_{ij} = k(g_j, t_i)$ were obtained from simulated trajectories (generated by FFPopSim) by binomial sampling (with size n_i) at specified time points t_i. Trajectory generation and sampling are implemented in the

Table 1 | Format of input data: the escape mutations are ordered first by the time of first observation and then by abundance.

Time (days)	Founder	Env581	Env830	Nef105	Gag236
9	5	2	0	0	0
45	0	0	5	3	0
85	0	0	0	0	8

Each entry in the table in a particular column reports the number of times a sequence is observed containing the escape of that column and all previous escape mutations.

file `model_fit/ctlutils.py` at http://git.tuebingen.mpg.de/ctlfit; see below.

Sequence data

The HIV sequences for patients CH40, CH58, and CH77 where downloaded from http://www.hiv.lanl.gov/content/sequence/HIV/USER_ALIGNMENTS/Salazar.html (8).

INFERENCE

The inference procedure consists of *initial guessing*, *sequential addition of escapes*, *multi-dimensional refinement*, and estimation of *posterior distributions*. The implementation can be found in `src/ctl_fit.py`, with the C code for the likelihood calculation in `src/cfit.cpp`.

Initial guesses

We produce initial guesses by single epitope modeling. The frequency of each escape mutation, ν_j, grows logistically with the escape rate (11). We expect that only the frequency of the first escape mutation is significantly affected by mutational input, because it receives input from the abundant founder sequence, whereas the later escapes only receive mutational input from the previously escaped genotype, which is still rare when the novel escape arises. Hence we only model the mutational dynamics of the first escape. In a single epitope model, the frequency of the founder variant is one minus the frequency of the escape variant. The frequency of the escape variant increases by $\mu(1-\nu_1)$ per day due to mutations from the founder and decreases by $\mu\nu_1$ due to further mutations to additional escapes. Combined with the logistic growth, the dynamics of ν_1 is described by

$$\dot{\nu}_1(t) = \varepsilon_1\nu_1(1-\nu_1) + \mu[1-2\nu_1]. \tag{6}$$

with initial condition $\nu_1(0) = 0$. Note the difference between the allele frequency ν, which refers to a particular escape mutation, and γ, which corresponds to frequencies of particular multi-epitope genotypes. The above ODE has the solution

$$\nu_1(t) = \frac{1}{2\varepsilon_1}\left[\varepsilon_1 - 2\mu + R\tanh\left(\frac{\alpha+t}{2}R\right)\right] \tag{7}$$

where $R = \sqrt{\varepsilon_1^2 + 4\mu^2}$ and $\alpha = \frac{4\mu-2\varepsilon_1}{4\mu^2+\varepsilon_1^2}$ (11). The escape rate ε_1 is determined by maximizing the likelihood (equation (5)) using `fmin` from `scipy` (43).

The seed time τ_j of subsequent escape mutants g_j is determined by maximizing the seed time prior $Q(\tau_j|\gamma_{j-1})$ defined in equation (4) using the previously determined γ_{j-1}. The frequencies of mutations are assumed to follow a logistic trajectory since the genotype from which they receive mutational input is itself still at low frequency:

$$\nu_j(t) = \frac{e^{\varepsilon_j(t-\tau_j)}}{e^{\varepsilon_j(t-\tau_j)} + N\varepsilon_j} \quad j > 1. \tag{8}$$

Again, we maximize the posterior probability, equation (5), to obtain an initial estimate of ε_j.

Sequential addition of escapes

Given the initial estimates for the first escape, we now add subsequent escapes to the multi-epitope model, which is formulated in terms of genotype counts k_{ij} and frequencies $\gamma_j(t)$. Note that the interpretation of genotype counts depends on how many epitopes are modeled. For example, if we model epitopes $1, \ldots, j$ out of a total of L epitopes, counts for genotype j are $k_{ij} = \sum_{l=j}^{L} k_{il}$, i.e., we ignore all later escapes.

If the added escape is unique, i.e., no other escape mutation has the exact same temporal pattern, we calculate the likelihood on a 21×31 grid of escape rates and seed times; comp. **Figure 4**. The grid spans values between 0 and twice the initial estimate for both the seed time and the escape rate. The most likely combination of seed time and escape rate is chosen, and the procedure is repeated with the next epitope.

If multiple epitopes exhibit the same temporal pattern, we add them all at once, constrain their escape rates to be equal, and assume they emerged in the order listed in the genotype matrix. Since we now have to optimize one joint escape rate and multiple seed times, we do not map the likelihood surface exhaustively but rather perform a greedy search. We examine next-neighbor moves with steps $\delta\tau = \pm 1$ day and $\delta\varepsilon = \pm 0.02$ per day, moves which change all seed times by $\delta\tau$, and 20 moves in which all seed times and escape rates are changed by $\delta\tau$ and $\delta\varepsilon$ with random sign; the step that maximizes the likelihood is accepted. This is repeated until no favorable move is found and further repeated with $\delta\varepsilon = \pm 0.01$ and ± 0.001 per day.

Refinement

We then iterate sequentially over every epitope and optimize its seed time and escape rate as described above, but with all other epitopes part of the multi-epitope model. This typically leads to rather small adjustments and converges rapidly.

Posterior distributions

To determine the posterior distribution of the escape rates, we attempt to change all seed times and escape rates by $\delta\tau = \pm 1$ day and $\delta\varepsilon = \pm 0.01$ per day with random sign. The move is accepted with probability $\min(1, \exp(\Delta))$, where Δ is the difference in log-likelihood before and after the change. We sample this Markov chain every 1000 moves and thereby map the posterior distribution of seed times and escape rates.

USAGE

All source code and scripts are available at http://git.tuebingen.mpg.de/ctlfit.

Building

The part of our method that is implemented in C and the python bindings can be built using `make` and the Makefile provided in the `src` directory. Prerequisites for building are `python2.7`, `scipy`, `numpy`, `swig`, and a `gcc` compiler.

Fitting

Given a text file with genotype counts specified as shown in **Table 1**, fitting is performed by calling the script `fit_escapes.py` with Python. Parameters can be set via command line arguments:

```
python fit_escapes.py --input datafile
```
(9)

where `--input` specifies the file with the genotype counts. Other parameters can be modified in a similar manner. Running the script with the option `--help` prints a list of all parameters. The estimated escape rates and seed times as well as the sampled posterior distribution will be saved in the directory `fit_escapes_output`, unless otherwise specified.

ACKNOWLEDGMENTS

We are grateful for stimulating discussions with F. Zanini. This work is supported by the ERC starting Grant HIVEVO 260686 (Richard A. Neher) and in part by the National Science Foundation under Grant No. NSF PHY11-25915. This work was performed under the auspices of the US Department of Energy under contract DE-AC52-06NA25396 and supported by NIH grants R01-AI028433, UM1-AI100645-01, and the National Center for Research Resources and the Office of Research Infrastructure Programs (ORIP) through grant R01-OD011095 (Alan S. Perelson).

REFERENCES

1. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. The immune response during acute HIV-1 infection: clues for vaccine development. *Nat Rev Immunol* (2009) **10**(1):11. doi:10.1038/nri2674
2. Fernandez CS, Stratov I, De Rose R, Walsh K, Dale CJ, Smith MZ, et al. Rapid viral escape at an immunodominant simian-human immunodeficiency virus cytotoxic T-lymphocyte epitope exacts a dramatic fitness cost. *J Virol* (2005) **79**(9):5721–31. doi:10.1128/JVI.79.9.5721-5731.2005
3. Li B, Gladden AD, Altfeld M, Kaldor JM, Cooper DA, Kelleher AD, et al. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J Virol* (2007) **81**(1):193–201. doi:10.1128/JVI.01231-06
4. Ganusov VV, Goonetilleke N, Liu MKP, Ferrari G, Shaw GM, McMichael AJ, et al. Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV infection. *J Virol* (2011) **85**(20):10518–28. doi:10.1128/JVI.00655-11
5. Seki S, Matano T. CTL escape and viral fitness in HIV/SIV infection. *Front Microbiol* (2012) **2**:267. doi:10.3389/fmicb.2011.00267
6. Asquith B, Edwards CTT, Lipsitch M, McLean AR. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol* (2006) **4**(4):e90. doi:10.1371/journal.pbio.0040090
7. Goonetilleke N, Liu MKP, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Ganusov VV, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med* (2009) **206**(6):1253–72. doi:10.1084/jem.20090365
8. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* (2009) **206**(6):1273–89. doi:10.1084/jem.20090378
9. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* (1996) **271**(5255):1582–6. doi:10.1126/science.271.5255.1582
10. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* (1995) **69**(8):5087–94.
11. Ganusov VV, Neher RA, Perelson AS. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. *J Stat Mech* (2013) **2013**(1):P01010. doi:10.1088/1742-5468/2013/01/P01010
12. Leviyang S. Computational inference methods for selective sweeps arising in acute HIV infection. *Genetics* (2013) **194**:737–52. doi:10.1534/genetics.113.150862
13. Neher RA, Leitner T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* (2010) **6**(1):e1000660. doi:10.1371/journal.pcbi.1000660
14. Batorsky R, Kearney MF, Palmer SE, Maldarelli F, Rouzine IM, Coffin JM. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc Natl Acad Sci U S A* (2011) **108**(14):5661–6. doi:10.1073/pnas.1102036108
15. da Silva J. The dynamics of HIV-1 adaptation in early infection. *Genetics* (2012) **190**(3):1087–99. doi:10.1534/genetics.111.136366
16. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* (2008) **105**(21):7552–7. doi:10.1073/pnas.0802203105
17. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, et al. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* (2010) **5**(8):e12303. doi:10.1371/journal.pone.0012303
18. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* (2012) **8**(3):e1002529. doi:10.1371/journal.ppat.1002529
19. Petravic J, Loh L, Kent SJ, Davenport MP. CD4+ target cell availability determines the dynamics of immune escape and reversion in vivo. *J Virol* (2008) **82**(8):4091–101. doi:10.1128/JVI.02552-07
20. Ganusov VV, De Boer RJ. Estimating costs and benefits of CTL escape mutations in SIV/HIV infection. *PLoS Comput Biol* (2006) **2**(3):e24. doi:10.1371/journal.pcbi.0020024
21. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. Dynamics of HIV-1 recombination in its natural target cells. *Proc Natl Acad Sci U S A* (2004) **101**(12):4204–9. doi:10.1073/pnas.0306764101
22. Zanini F, Neher RA. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics* (2012) **28**(24):3332–3. doi:10.1093/bioinformatics/bts633
23. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, Perelson AS, et al. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J Virol* (2003) **77**(8):5037–8. doi:10.1128/JVI.77.8.5037-5038.2003
24. Sunnker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian computation. *PLoS Comput Biol* (2013) **9**(1):e1002803. doi:10.1371/journal.pcbi.1002803
25. Basu D. On the elimination of nuisance parameters. *J Am Stat Assoc* (1977) **72**(358):355–66. doi:10.1080/01621459.1977.10481002
26. Kepler TB, Perelson AS. Modeling and optimization of populations subject to time-dependent mutation. *Proc Natl Acad Sci U S A* (1995) **92**(18):8219–23. doi:10.1073/pnas.92.18.8219
27. Desai MM, Fisher DS. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* (2007) **176**(3):1759–98. doi:10.1534/genetics.106.067678
28. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* (1995) **267**(5197):483–9. doi:10.1126/science.7824947

29. Perelson AS, Essunger P, Ho DD. Dynamics of HIV-1 and CD4+ lymphocytes in vivo. *AIDS* (1997) **11**(Suppl A):S17–24.
30. Read EL, Tovo-Dwyer AA, Chakraborty AK. Stochastic effects are important in intrahost HIV evolution even when viral loads are high. *Proc Natl Acad Sci U S A* (2012) **109**(48):19727–32. doi:10.1073/pnas.1206940109
31. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. *Theor Popul Biol* (2009) **75**(4):286–300. doi:10.1016/j.tpb.2009.02.006
32. Neher RA, Shraiman BI. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* (2011) **188**:975–96. doi:10.1534/genetics.111.128876
33. Haase AT, Henry K, Zupancic M, Sedgewick G, Faust RA, Melroe H, et al. Quantitative image analysis of HIV-1 infection in lymphoid tissue. *Science* (1996) **274**(5289):985–9. doi:10.1126/science.274.5289.985
34. Liu S-L, Mittler JE, Nickle DC, Mulvania TM, Shriner D, Rodrigo AG, et al. Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. *J Virol* (2002) **76**(21):10674–84. doi:10.1128/JVI.76.21.10674-10684.2002
35. Paredes R, Sagar M, Marconi VC, Hoh R, Martin JN, Parkin NT, et al. In vivo fitness cost of the m184v mutation in multidrug-resistant human immunodeficiency virus type 1 in the absence of lamivudine. *J Virol* (2009) **83**(4):2038–43. doi:10.1128/JVI.02154-08
36. Bonhoeffer S, Barbour AD, De Boer RJ. Procedures for reliable estimation of viral fitness from time-series data. *Proc Biol Sci* (2002) **269**(1503):1887–93. doi:10.1098/rspb.2002.2097
37. Asquith B, McLean AR. In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proc Natl Acad Sci U S A* (2007) **104**(15):6365–70. doi:10.1073/pnas.0700666104
38. Petravic J, Ribeiro RM, Casimiro DR, Mattapallil JJ, Roederer M, Shiver JW, et al. Estimating the impact of vaccination on acute simian-human immunodeficiency virus/simian immunodeficiency virus infections. *J Virol* (2008) **82**(23):11589–98. doi:10.1128/JVI.01596-08
39. Josefsson L, King MS, Makitalo B, Brännström J, Shao W, Maldarelli F, et al. Majority of CD4+ T cells from peripheral blood of HIV-1-infected individuals contain only one HIV DNA molecule. *Proc Natl Acad Sci U S A* (2011) **108**(27):11199–204. doi:10.1073/pnas.1107729108
40. Illingworth CJR, Mustonen V. A method to infer positive selection from marker dynamics in an asexual population. *Bioinformatics* (2012) **28**(6):831–7. doi:10.1093/bioinformatics/btr722
41. Boltz VF, Ambrose Z, Kearney MF, Shao W, Ramani VNK, Maldarelli F, et al. Ultrasensitive allele-specific PCR reveals rare preexisting drug-resistant variants and a large replicating virus population in macaques infected with a simian immunodeficiency virus containing human immunodeficiency virus reverse transcriptase. *J Virol* (2012) **86**(23):12525–30. doi:10.1128/JVI.01963-12
42. Liu MKP, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J Clin Invest* (2013) **123**(1):380–93. doi:10.1172/JCI65330
43. Oliphant T. Python for scientific computing. *Comput Sci Eng* (2007) **9**(3):10–20. doi:10.1109/MCSE.2007.58

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2013; paper pending published: 15 June 2013; accepted: 12 August 2013; published online: 03 September 2013.

*Citation: Kessinger TA, Perelson AS and Neher RA (2013) Inferring HIV escape rates from multi-locus genotype data. Front. Immunol. **4**:252. doi: 10.3389/fimmu.2013.00252*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2013 Kessinger, Perelson and Neher. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 30 January 2014
doi: 10.3389/fimmu.2014.00020

Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus

Paola Carrillo-Bustamante*, Can Keşmir and Rob J. de Boer

Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, Netherlands

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Becca Asquith, Imperial College London, UK
Jayajit Das, The Ohio State University, USA

****Correspondence:***
Paola Carrillo-Bustamante, Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Padualaan 8, Utrecht 3584 CH, Netherlands
e-mail: p.carrillobustamante@uu.nl

The responsiveness of natural killer (NK) cells is controlled by balancing signals from activating and inhibitory receptors. The most important ligands of inhibitory NK cell receptors are the highly polymorphic major histocompatibility complex (MHC) class I molecules, which allow NK cells to screen the cellular health of target cells. Although these inhibitory receptor–ligand interactions have been well characterized, the ligands for most activating receptors are still unknown. The mouse cytomegalovirus (MCMV) represents a helpful model to study NK cell-driven immune responses. Many studies have demonstrated that CMV infection can be controlled by NK cells via their activating receptors, but the exact contribution of the different signaling potential (i.e., activating vs. inhibiting) remains puzzling. In this study, we have developed a probabilistic model, which predicts the optimal specificity of inhibitory and activating NK cell receptors needed to offer the best protection against a CMV-like virus. We confirm our analytical predictions with an agent-based model of an evolving host population. Our analysis quantifies the degree of protection of each receptor type, revealing that mixed haplotypes (i.e., haplotypes composed of activating and inhibiting receptors) are most protective against CMV-like viruses, and that the protective effect depends on the number of MHC loci per individual.

Keywords: NK cell receptors, evolution, CMV infection, models, theoretical, agent-based modeling

INTRODUCTION

Natural killer (NK) cells contribute to the host immune response by recognizing and killing viral-infected and tumor cells (1). Their activity is controlled by balancing signals from a vast repertoire of activating and inhibiting receptors enabling them to distinguish healthy from unhealthy cells (2). The most important ligands for inhibitory NK cell receptors (iNKR) are MHC class I molecules on other cells. An infected cell may have lower MHC expression, altering the binding with inhibitory receptors, disrupting the balance of signals, and allowing for NK cell activation. The mechanism by which NK cells attack MHC class I deficient cells was coined by Kärre et al. as "missing-self" detection (3).

There are several NKRs that contribute to missing-self detection. In humans, for example, the inhibitory receptor CD96/NKG2A binds to complexes of the human leukocyte antigen-E (HLA-E), which presents peptides derived from the leader sequences of HLA-A, -B, and -C molecules (4, 5). Both the receptor and the ligand are highly conserved in these inhibitory interactions, and the down-stream effects are remarkably similar across individuals (6). The killer immunoglobulin-like receptors (KIRs) also contribute to monitor abnormalities in MHC class I expression on cell surfaces. In contrast to the CD96/NKG2 superfamily, KIRs are highly polygenic and polymorphic, exhibit both inhibitory and activating potential, and bind to the highly polymorphic HLA-A, -B, and -C molecules (7–9). Consequently, the interactions between KIRs and classical HLA-class I molecules are very diverse (10). Thus, humans have two types of NKRs, one conserved and one highly diverse, performing seemingly the same function.

Humans are not the only species that have an expanded and polymorphic NKR gene complex. During mammalian radiation, many different species have diversified alternative NKR gene families recognizing MHC class I. This example of convergent evolution includes three gene families from two structurally unrelated superfamilies: KIRs, the CD94/NKG2, and the Ly49 (11). Higher primates have expanded their KIR genes (12); a group of lower primates have expanded NKG2 (13), whereas rodents and equids have expanded Ly49 (14, 15). These alternative genetic strategies illustrate the evolutionary complexity of these systems, and suggest that an expanded NKR gene complex is beneficial for survival. But, if conserved inhibitory receptor–ligand interactions (such as NKG2A–HLA-E in humans) are capable to successfully detect missing-self, why have several NKR families evolved to become polygenic and polymorphic? Even more intriguing, why have they evolved receptors with activating potential?

In humans, some activating NKRs (aNKRs) are associated with the disease outcome of viral infections and malignancies (16). For example, in combination with HLA–Bw4, the activating KIR3DS1 has been associated with a delayed progression to AIDS in HIV-1 infected individuals (17, 18). KIR3DS1 has also been linked to an increased rate of spontaneous recovery after hepatitis B infections (19), a reduced risk of developing hepatocellular carcinoma in patients infected with HCV (20), and a reduced risk of Hodgkin's lymphoma (21). Moreover, maternal activating KIRs are related to protection against several pregnancy disorders (22). But because only a few ligands for activating KIRs have been identified so far, the exact mechanisms underlying the provided protection in humans remain puzzling.

Studies in mice have revealed important insights into the role of aNKR during viral infections (23, 24). Viruses like the mouse cytomegalovirus (MCMV) down-regulate the expression of MHC class I molecules from the cell surface to escape T cell response, and may additionally code decoy MHC molecules (m157) that can inhibit NK cell activation (23). Mouse strains that are resistant to MCMV carry the activating Ly49H gene, which binds with high affinity to the MHC-like viral protein m157. In contrast, mice susceptible to MCMV lack the activating gene but carry the inhibiting receptor Ly49I, which also binds strongly to the m157 protein. The activating Ly49H emerged from an inhibitory counterpart (25), suggesting that the evolution of an aNKR was due to the immune pressure induced by the "MHC decoy" m157 during CMV infection (26, 27).

Although these studies shed light into the importance of NKR in general, the specific contribution of activating and inhibitory receptors to the NK cell response is still unknown. We previously studied the evolution of KIR diversity in a human population infected with CMV-like viruses by using a computational agent-based model (28). We showed that iNKRs require sufficient specificity to protect populations against viruses evolving MHC-like molecules, and that diversity in the NK cell genetic complex evolves as a result of the required discrimination between self-MHC molecules and viral decoy molecules. Here, we also consider aNKRs, and develop a probabilistic model to quantify the optimal specificity of inhibitory and activating NKRs needed to render maximal protection against CMV-like viruses. We also analyze the effect of mixed haplotypes (i.e., composed of aNKR and iNKR) on protection, and confirm the expectations of the probabilistic model with an agent-based computational model. Our studies reveal that mixed haplotypes composed of specific activating and inhibitory NKRs render high protection against CMV-like viruses encoding for decoy molecules, and that the protective effect depends on the number of MHC loci per individual.

RESULTS

We analyze the effect of the specificity of activating and inhibitory NKRs on the detection of a virus presenting MHC-like molecules with a simple probabilistic model. Our model estimates the chance of protection P, i.e., the probability of a host detecting an infection by NK cells, as a function of the haplotype size, specificity (i.e., the probability p of recognizing any random MHC molecule), and number of MHC loci.

The responsiveness of NK cells (i.e., their ability to discriminate cells with normal MHC expression from those lacking MHC) is regulated by a process called "education" or "licensing" taking place during NK cell development (29). During this process, the interactions of iNKRs with their MHC ligands render the NK cells with functional competence (13, 29, 30). To prevent NK cell-related autoimmunity, activating receptors also participate in the education process, where the chronic exposure of aNKR ligands during development results in hyporesponsive NK cells (31, 32).

For simplicity, we do not model individual NK cells, each expressing a random set of tuned receptors. We rather consider for each individual a global repertoire of receptors, which have the potential to license NK cells. Henceforth, we will refer to these receptors as "licensed" receptors. We mimic the MHC-dependent education process during NK cell development by creating a repertoire of NKRs composed of iNKRs that recognize at least one of the MHC molecules of the host, and of aNKRs that recognize none of the MHC molecules of the host (**Figure 1**). By considering a global repertoire, we assume that there will be at least one subset of NK cells expressing at least one of the "licensed" NKRs. Upon infection, we consider only those NK cell subsets having licensed receptors. If these can successfully detect the virus, they will become activated, expand, and protect (see Appendix for a full discussion). Therefore, only the licensed repertoire of NKRs is allowed to participate in the immune response.

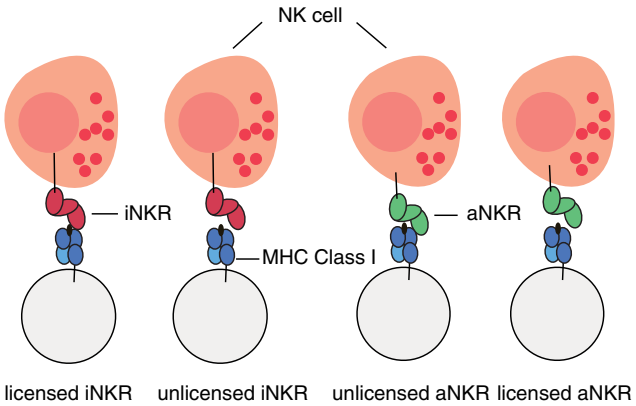


FIGURE 1 | Cartoon of NK cell education. iNKRs (represented in red) recognizing at least one of the MHC molecules per individual will become licensed. In contrast, aNKRs (depicted in green), which do not recognize *any* of the host's MHC molecules will become licensed.

Whether a decoy protein allows a virus to successfully escape the NK response, i.e., whether that individual is protected against the infection, depends on the receptor type and on the receptor specificity. iNKRs that bind the decoy molecule cannot detect missing-self and are "fooled" by the decoy. Conversely, aNKRs binding the foreign decoy protein can specifically recognize the infection and therefore protect the host. With this model, we quantify the contribution of each receptor type and its specificity to the detection of CMV-like viruses.

INHIBITORY AND ACTIVATING NK CELL RECEPTORS DIFFER IN THE PROTECTION LEVEL THEY PROVIDE

There are two crucial processes for a single iNKR to detect virus evolving MHC-like molecules. First, iNKRs have to be licensed during the NK cell education to become fully functional during an immune response. Second, iNKRs should not bind decoy molecules upon infection. Because, in our model, iNKRs are only licensed if they recognize at least one of the MHC molecules in their host, and decoy molecules are similar to self-MHC molecules, iNKRs face the challenge of distinguishing self-MHC molecules from foreign decoy molecules. We previously demonstrated that this challenge can be solved by evolving sufficiently specific iNKRs (28). In that study, we defined specificity as the probability (p) of any NKR to recognize a random MHC molecule in the population. Herewith, degenerate receptors (i.e., with $p = 1$) are able to

recognize all MHC molecules in the population, whereas specific receptors (i.e., with $p \sim 0$) recognize only a small fraction of them. Since the exact relation between ligand–receptor binding affinity and signaling potential remains unknown, we do not consider different binding affinities here, and we model discrete MHC–NKR interactions.

To study whether there is an optimal specificity for which iNKRs are not inhibited by such "decoy viruses," we calculated the probability of licensed iNKRs detecting the infection. A single iNKR becomes licensed with a probability $q_{\mathrm{I}} = 1-(1-p_{\mathrm{I}})^{2N_{\mathrm{MHC}}}$, where p_{I} describes the specificity (i.e., the probability of any iNKR to recognize any MHC in the population), and N_{MHC} the number of MHC loci per individual. The probability of a haplotype composed of N_{iNKR} to have exactly licensed iNKRs is given by the binomial distribution as follows:

$$P\left(\mathrm{iNKR}_{\mathrm{licensed}} = \ell\right) = \binom{N_{\mathrm{iNKR}}}{\ell} \left(1-q_{\mathrm{I}}\right)^{N_{\mathrm{iNKR}}-\ell} q_{\mathrm{I}}^{\ell}. \quad (1)$$

To successfully detect a decoy virus, none of the licensed iNKRs should bind the decoy molecule. Thus, the overall probability of detecting the infection is determined by the chance that none of the licensed iNKRs recognizes a decoy molecule, and can be calculated by:

$$P_{\mathrm{I}}\left(\mathrm{detection}\right) = \sum_{\ell=1}^{N_{\mathrm{iNKR}}} \binom{N_{\mathrm{iNKR}}}{\ell} \left(1-p_{\mathrm{I}}\right)^{\ell} \left(1-q_{\mathrm{I}}\right)^{N_{\mathrm{iNKR}}-\ell} q_{\mathrm{I}}^{\ell}. \quad (2)$$

Our analysis confirms that for any haplotype size, there is an optimal specificity. For $N_{\mathrm{iNKR}} \leq 25$, our model predicts a maximal level of protection (i.e., $P_{\mathrm{I}} = 0.85$), which can only be obtained with high specificity values ($p_{\mathrm{I}} \leq 0.2$) and a large number of genes per haplotype ($N_{\mathrm{iNKR}} \geq 20$) (**Figure 2A**).

A host with degenerate iNKRs (e.g., $p_{\mathrm{I}} \geq 0.8$) has a large repertoire of licensed iNKRs. But because of the low specificity, the iNKRs within that individual are expected to also recognize any foreign decoy molecule as self, offering no protection. In contrast, when iNKRs are specific (e.g., $p_{\mathrm{I}} \leq 0.2$) the repertoire of licensed iNKRs per individual is lower, but if there are several genes per haplotype, the chance of having at least one licensed specific iNKR increases. Due to their high specificity, it is unlikely for a licensed iNKR to also recognize a foreign decoy molecule, impeding the virus to escape the NK immune response. Therefore, an infection with a decoy virus can be controlled with a probability of at least 70% in a haplotype composed of more than 10 iNKRs when $p_{\mathrm{I}} \leq 0.25$. Thus, the probability of detecting the virus increases with both a higher specificity and a larger number of genes per haplotype (i.e., N_{iNKR}). This confirms our previous results, suggesting that large haplotypes composed of non-overlapping specific iNKRs are most protective (28).

We next developed a model considering only aNKRs. Similar to the iNKRs, the two crucial processes for an aNKR to detect the virus depends on the probability of becoming licensed and recognizing the decoy molecule as a foreign antigen. However, the licensing process is almost opposite between aNKRs and iNKRs. An aNKR becomes licensed if it does not recognize any MHC molecule within an individual. The probability of a single aNKR to become licensed is therefore described by $q_{\mathrm{A}} = (1-p_A)^{2N_{\mathrm{MHC}}}$, where p_{A} is the specificity of an aNKR. Opposite to an iNKR, an aNKR detects a "decoy virus" if it binds the MHC decoy. Thus, the overall probability of protection in this case is determined by the chance of at least one licensed aNKR binding the decoy molecule, and is given by:

$$P_{\mathrm{A}}\left(\mathrm{detection}\right) = \sum_{\ell=1}^{N_{\mathrm{aNKR}}} \binom{N_{\mathrm{aNKR}}}{\ell} \left(1-\left(1-p_{\mathrm{A}}\right)^{\ell}\right) \times \left(1-q_{\mathrm{A}}\right)^{N_{\mathrm{aNKR}}-\ell} q_{\mathrm{A}}^{\ell}, \quad (3)$$

where N_{aNKR} is the number of aNKRs per haplotype.

This model reveals that there is again an optimal specificity, and the protection range for aNKR is much broader than that for iNKR, covering also less specific receptors (i.e., $0.1 \leq p_{\mathrm{A}} \leq 0.7$) (**Figure 2B**). In these cases, the optimal protection (i.e., $P_{\mathrm{A}} = 1$) is obtained with haplotypes composed of 12 genes, having intermediate specificity values ($0.2 \leq p_{\mathrm{A}} \leq 0.65$). To avoid self-reactivity, aNKRs become licensed only if they fail to recognize all self-MHC molecules. Additionally, an aNKR must recognize foreign MHC-like molecules to detect the infection. Therefore, the challenge for an aNKR is opposite to that of an iNKR, since it must recognize foreign antigens but not self-MHC molecules. A degenerate aNKR will recognize every decoy in the population but it will never become licensed. As a result, the optimal protection is reached in large haplotypes composed of aNKRs with intermediate specificity.

Note that we consider individuals to be heterozygous for all MHC loci. Allowing individuals to be homozygous in some MHC loci does not qualitatively change our results on specificity and protection, since MHC homozygosity has only a mild effect on the number of licensed receptors, ℓ (results not shown).

THE PROTECTION LEVEL DEPENDS ON THE NUMBER OF MHC LOCI

Above, we considered only one MHC locus per individual as a representation of HLA-C as the main identified ligand for inhibitory KIRs. However, HLA-A and -B molecules have also been identified as KIR ligands, and HLA-E is the ligand for CD94/NKG2A. Therefore, we expanded our model to consider two MHC loci per individual. The distribution of protection levels is similar to the model with one MHC locus, showing a small protective area for individuals carrying only iNKRs (**Figure 2C**), whereas individuals carrying aNKRs have a broader protective range (**Figure 2D**). However, the area of maximal protection is skewed in both cases. Because iNKRs have to recognize at least one self-MHC molecule to become licensed, the chance of having several licensed NKRs per haplotype increases by having 2 MHC loci (and thus 4 MHC molecules per heterozygous individual). Therefore, a high protection (e.g., $P_{\mathrm{I}} \geq 0.85$) can be reached already with a smaller haplotype, e.g., one composed of at least 11 iNKRs.

In contrast, the probability of an aNKR to become licensed decreases with 2 MHC loci because aNKRs should not recognize *any* of the MHC molecules within an individual. Consequently, the protection with aNKRs reaches high values (i.e., $P_{\mathrm{A}} \geq 0.85$) only with large haplotypes composed of at least 20 genes and the optimal protection level ($P_{\mathrm{A}} = 1$) is never obtained. Thus, the

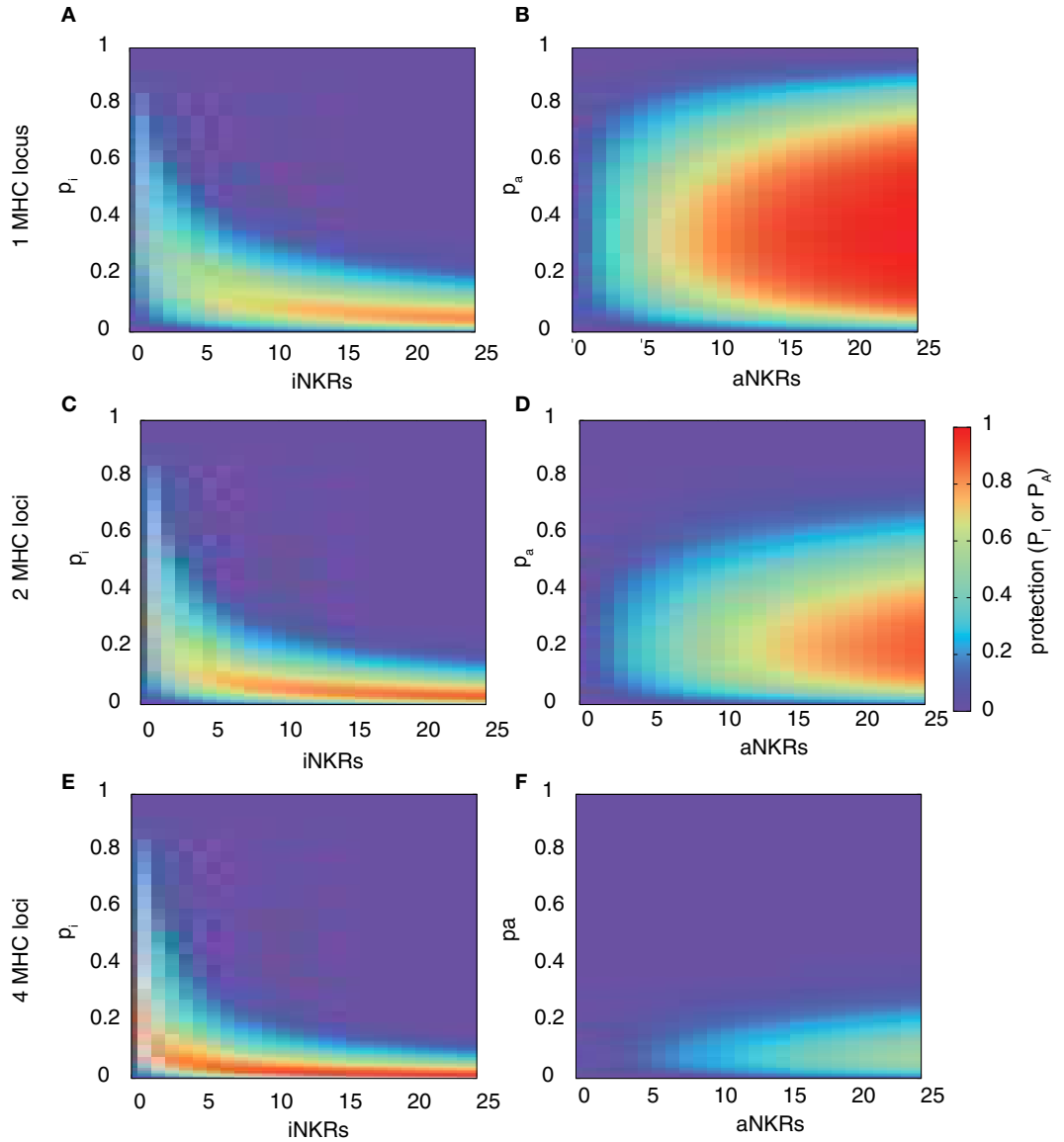


FIGURE 2 | Range of protection differs between iNKRs and aNKRs. The heatmaps show the protection level as the probability of detecting the infection with a virus expressing a decoy molecule to mask MHC down-regulation. In the left column, the protection for individuals carrying only iNKRs [calculated by equation (2)] is shown, whereas in the right column, the protection for individuals carrying only aNKRs [calculated by equation (3)] is depicted. The protection level is shown in the color bar from highest (red) to lowest (blue). p_i and p_a correspond to the specificity of iNKR and aNKR, respectively. **(A,B)** The protective range for iNKRs is small and skewed toward a large haplotype size and high specific values. In contrast, aNKRs offer a broad range of protection for intermediate specificity values and a smaller haplotype size. Calculations were done with 1 MHC locus **(A,B)**, 2 MHC loci **(C,D)**, and 4 MHC loci **(E,F)**.

protection of aNKRs is highly dependent on the number of MHC molecules per individual.

With even higher MHC complexity, i.e., by increasing the number of MHC loci per individual to 4, fewer iNKRs are sufficient to successfully clear the infection (**Figures 2E,F**). Because of the education process in our model, hosts with 4 MHC loci have a much larger licensed iNKR repertoire compared to individuals having 1 MHC locus. These hosts reach the maximal protection already with a haplotype size of 4 receptors. Even for lower haplotype sizes, a good protection level (i.e., $0.3 \leq P_I \leq 0.7$) can be reached at lower specificity values ($p_I \leq 0.35$) (**Figure 2E**). This effect was further increased when considering 8 MHC loci per individual, where the maximal protection was reached with only one specific iKIR (results not shown).

However, an expanded MHC haplotype is disadvantageous for individuals having only aNKRs. Because in our model the licensing process is more difficult with a higher number of MHC molecules, little protection can be provided. The infection can be controlled

with a maximal probability of 50 and 35% in individuals with 4 (**Figure 2F**) and 8 MHC loci (results not shown), respectively.

Taken together, these results show that aNKRs provide little protection against a virus evolving MHC decoy proteins in individuals having several MHC loci, and that a contracted haplotype of iNKRs is already protective when the MHC complexity increases.

VIRAL DETECTION IS MAXIMAL IN MIXED HAPLOTYPES

To predict the combined protection of activating and inhibitory NKRs, we expanded our model and considered mixed haplotypes, i.e., haplotypes composed of both iNKRs and aNKRs. We predict the combined probability of detecting the virus as follows:

$$P = 1 - (1 - P_{\mathrm{I}})\,(1 - P_{\mathrm{A}})\,. \tag{4}$$

We computed the protection in hosts carrying two MHC and 20 NKR loci, and varied the fraction of aNKRs in the NKR haplotype, while keeping the total number of loci constant. The best protection is reached in mixed haplotypes (**Figure 3**). As seen above, haplotypes with aNKRs only provide protection (i.e., $0.5 \leq P \leq 0.8$) for intermediate specificity values $0.15 \leq p_{\mathrm{A}} \leq 0.4$ (**Figure 3A**). With increasing number of iNKRs per haplotype, the protection

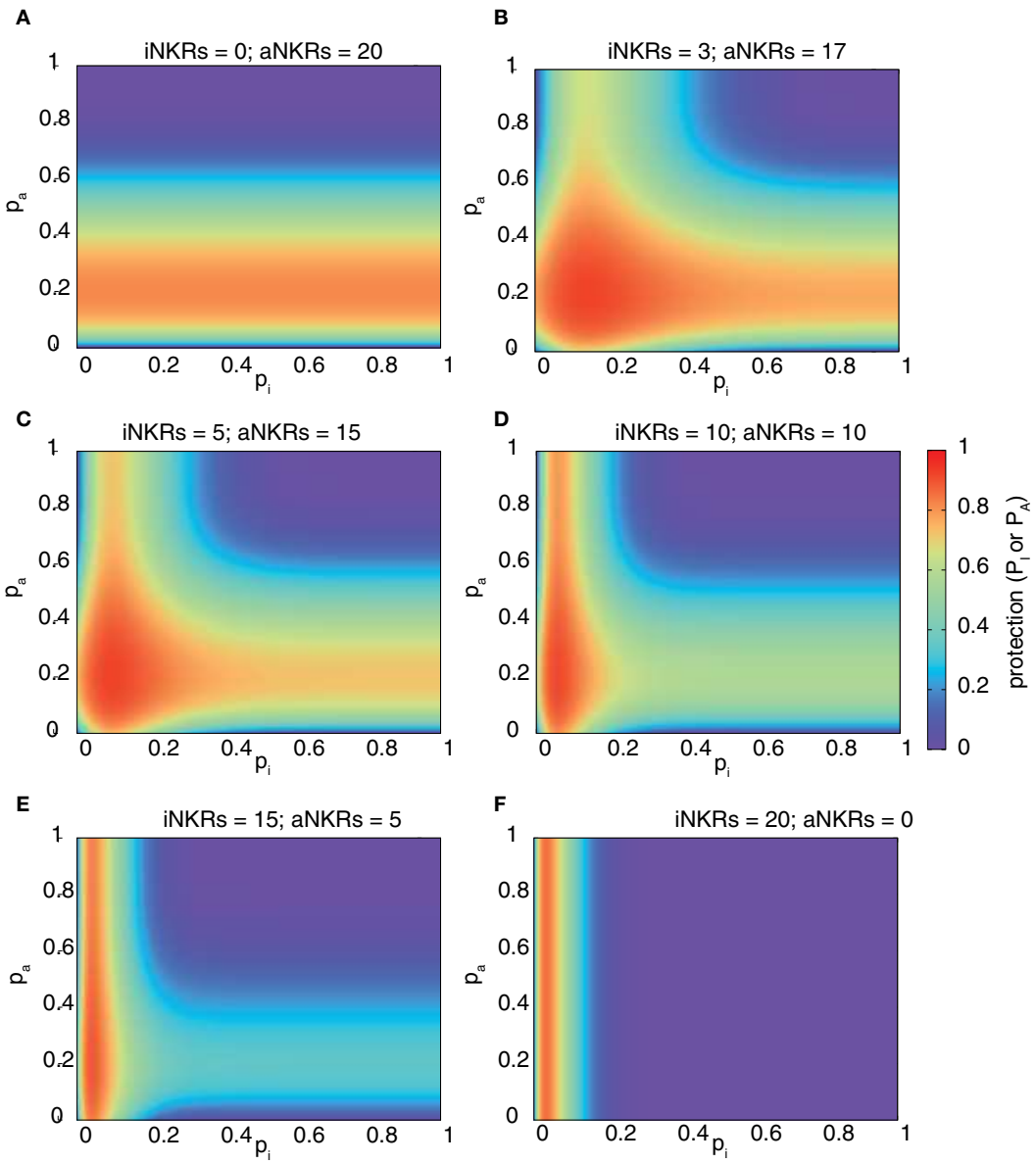


FIGURE 3 | Mixed haplotypes render highest protection. The heatmaps show the protection level as the probability of detecting the infection for haplotypes composed of iNKRs and aNKRs [as calculated by equation (4)]. The protection level is shown in the color bar from highest (red) to lowest (blue). **(A–F)** Show the NKR haplotype composition for different fractions of iNKR and aNKR. Considering a haplotype size of $N_{\mathrm{NKR}} = 20$, we first modeled haplotypes composed of aNKRs only **(A)**, and reduced their number, while increasing the number of iNKRs **(B–E)**, until we obtained a haplotype composed only of iNKRs **(F)**. We here consider 2 MHC loci (i.e., $N_{\mathrm{MHC}} = 2$).

reaches higher values (approaching $P = 1$) (**Figures 3B,C**), covering a larger range of specificity values and having a skewed distribution toward more specific inhibitory and activating receptors. A large number of iNKRs per haplotype reduces the contribution of aNKRs, and therefore the latter can have low specificity values without affecting the protection level (**Figures 3D,E**). Note that the area of high protection shrinks when the fraction of aNKRs is decreased, where maximal protection can only be achieved for extremely high specificities (i.e., $p_{\mathrm{I}} \leq 0.1$) (**Figure 3F**).

These results depend on a similar manner on the number of MHC loci per individual as those shown in **Figure 2**, with aNKRs having a lower protective effect with increasing MHC loci number (results not shown). Therefore, we conclude that the maximal protection against CMV-like viruses is easier to achieve in mixed haplotypes.

AGENT-BASED MODEL OF ACTIVATING AND INHIBITORY NKRs

Our probabilistic model allows us to quantify the expected protection, given a certain number of aNKRs and iNKRs. However, it is not clear whether a population with evolving NKRs would find the same basin of attraction for the specificity (i.e., p) when infected with a CMV-like virus.

To study the evolution of NKR specificity in populations infected with a CMV-like virus, we developed an agent-based model similar to the one published in Ref. (28) (for a detailed description of the model, see Materials and Methods). Briefly, our model considers a human population infected with a non-lethal herpes-like virus causing chronic infections. The hosts carry a diploid genome with one or two MHC loci and ten NKR loci. We consider 15 MHC alleles per locus (mimicking the common HLA-B and -C alleles in the European populations) and this polymorphism is kept constant throughout the entire simulation (i.e., we do not allow for mutation of the MHC genes). The initial NKR haplotype consists of ten different genes, and all individuals are homozygous for the same NKR haplotype.

Upon birth, novel receptors can be created, allowing for evolution within the NKR gene complex. Each new receptor comes with a randomly chosen receptor type (i.e., either inhibitory or activating) and a randomly chosen specificity value (corresponding to $0 \leq p \leq 1$, see Materials and Methods). Receptors are so specific that they are unable to recognize any MHC in the population will never be functional, and are considered to be pseudogenes. Thus, haplotypes expand by acquiring receptors with novel p values and signaling potential, but can also contract due to the accumulation of pseudogenes.

In this agent-based model, we also mimic the MHC-dependent NK cell education process (**Figure 1**). We remove iNKRs which fail to recognize any MHC molecule within an individual from the licensed repertoire. Similarly, those aNKRs capable of recognizing self-MHC molecules are deleted from the licensed repertoire. Only the licensed NKRs are able to participate during the immune response.

Infection of a host starts with a short acute phase, after which the individual either recovers or becomes chronically infected. We consider one wild-type virus and several decoy viruses (1 decoy per MHC molecule in the population). We do not allow for superinfection nor co-infection, thus hosts can be infected with only one of the viruses. A decoy virus down-regulates the expression of all MHC molecules within an individual, and expresses an MHC-like molecule. Thus, every virus expressing a decoy molecule has the potential to escape the immune response of both T and NK cells. The evolution of decoy proteins is modeled by allowing the virus to adopt a randomly selected MHC molecule from its host. Therefore, each decoy protein is actually an MHC molecule.

The population is first inoculated with the wild-type virus, which can be typically cleared after the acute phase because of the implicit response of both T and NK cells. We model the immune response with one parameter describing the probability of clearing the infection. For the wild-type virus, this is set to $p_{wt} = 0.85$ (**Table 1**), resulting in approximately 85% of the wild-type infections being cleared. Individuals clearing the infection become immune for a period t_{i} of 10 years. At steady state, approximately 20% of the population becomes chronically infected (**Figures 4A,B**; green solid lines), 65% become immune (**Figures 4A,B**; green dashed lines), and 5% are susceptible for infection. The immune escape of the decoy viruses is modeled by decreasing the clearance probability to zero ($p_{\mathrm{dec},1} = 0$, **Table 1**), which occurs if at least one of the licensed iNKRs or none of the aNKRs binds to the decoy molecule (**Table 2**). With this agent-based model, we can study the evolution of NKR specificity, and quantify the protection provided by activating and inhibitory receptors.

INHIBITORY RECEPTORS EVOLVE HIGHER SPECIFICITY THAN ACTIVATING RECEPTORS AFTER A CMV-LIKE INFECTION

We first study the protection provided by iNKRs against a CMV-like virus. After 5000 years of infection with the wild-type virus, we allow for the emergence of decoy viruses. The initial specificity of the iNKRs is set to $p \approx 0.4$ (see Materials and Methods). The decoy viruses spread easily among individuals carrying only iNKRs, resulting in a high fraction of chronically infected individuals (**Figure 4A**; red solid lines). Moreover, almost none of the hosts is able to control the infection (**Figure 4A**; red dashed lines), and the total population size decreases dramatically to 50% of the carrying capacity (i.e., maximal population size), confirming the results from the probabilistic model. However, after centuries of infection, the fraction of recovered individuals increases, and with it the total population size, indicating a recovery of the population. This observation is consistent in all ten simulations we performed for iNKRs (**Figures 4A,C**).

To study how these individuals evolve to control an infection with a virus having an MHC-like molecule, we analyze the average specificity of the NKRs over time. We determine how many MHC molecules in the population can be recognized by each receptor, and normalize it by the number of total MHC molecules in the population (**Figure 4E**; red line). We observe that the specificity increases after the emergence of the decoy viruses. At the end of the simulations, each iNKR recognize <20% of the MHC molecules in the population, indicating that evolution selects for specific iNKRs.

We perform the same simulations and analysis for populations having only aNKRs. Compared to populations having iNKRs, the initial spread of the virus is somewhat impaired (**Figure 4B**; red solid lines). Already at the beginning of the infection, some

Table 1 | Parameters of the agent-based model.

Parameter	Value
Time step	1 week
Simulation time	2 Million years
HOST PARAMETERS[a]	
Maximal population size, N_{max}	5000 Individuals
MHC diversity	1–2 Loci, each with 15 alleles
Number of NKR loci	5–10
Bit string length	16 Bits[b]
Host mutation rate, μ	0.00005 Per gene per birth event
INFECTION[c]	
Infection state, i	1 (Acute), 2 (chronic)
Effect of viral load on the death rate, VL_i	0.1 (For $i=1$), 0.06 (for $i=2$) per year
Probability of viral transmission during acute phase, p_{ac}	0.85 Per contact
Probability of viral transmission during chronic phase, p_{ch}	0.15 Per contact
Probability of clearing the wild-type virus, p_{wt}	0.85
Success state of the decoy virus, s	0 (Successful), 1 (unsuccessful)
Probability of clearing the virus evolving decoy molecules, $p_{dec,s}$	0 (For $s=0$), 0.5 (for $s=1$)
Immunity time, t_i	10 years
Acute infection time, t_{inf}	4 weeks
VIRUS PARAMETERS	
Virus mutation rate, μ_v[d]	0.0001 Per week
INITIAL CONDITIONS	
Initial population size, N_{init}	4500 Individuals
KIR initial diversity (SRI)	5–10 (1 Allele per locus)

[a]The death and birth rate parameters are age-dependent and have been chosen according to a human population (33). For a full description of the age-dependency of birth and death rate, see Ref. (28).

[b]The choice to use 16-bit strings represents a large enough theoretical repertoire of 65,536 sequences.

[c]The parameters used for the infection are chosen to maintain the epidemic. Changing the length of the acute phase or the probabilities of clearance do not affect our results on the evolution of the NKRs qualitatively (results not shown).

[d]We manually switch on the mutation of the viruses at specific points in time, and after that the mutation rate determines the waiting time for the mutant to arrive. The mutant viruses appear in a short time scale and once the virus has spread in the population, mutation does not occur anymore. Since we analyze the genetic diversity long after the arrival of the virus, changes in mutation rate should not affect the outcome.

individuals are able to control the virus (**Figure 4B**; red dashed lines). The population size decreases to 60% of the carrying capacity, and is therefore fitter than in those simulations considering only iNKRs (**Figures 4A–C**). Thus, aNKRs provide a better initial protection than iNKRs. Accordingly, the number of recovered individuals and thereby the total population increases rapidly, reflecting their fast recovery against viruses evolving decoy proteins.

The higher protection of aNKRs compared to iNKRs can be explained by the initial specificity. Because we initialize all populations with intermediate specificity, individuals carrying only aNKR are initially better protected (**Figure 2**). Nevertheless, aNKRs also evolve to be more specific (**Figure 4E**; black lines). At the end of the simulation, aNKRs recognize on average approximately 35% of all MHC molecules, and hence decoys in the population. Taken together, our agent-based model reveals that iNKRs need to be more specific than aNKRs to protect during an infection with a CMV-like virus, confirming the results from our probabilistic model.

Note that we do not explore all possible loci number in the agent-based model. To save computational time, we test the evolution of the specificity given a fixed loci number of NKRs. Populations carrying 10 NKR loci correspond to 20 NKRs in the probabilistic model, where the protection is maximal at very high specificity values for iNKR, and intermediate values for aNKR. These values correspond indeed to the specificity values that the populations evolve in our simulations. We carried out additional simulations for 5 and 15 NKR loci, the results of which confirmed the predictions of the mathematical model (results not shown).

POPULATIONS HAVING ONLY aNKRs EVOLVE A LARGER NKR POLYMORPHISM THAN POPULATIONS WITH ONLY iNKRs

Our probabilistic model predicts that the protection by iNKRs and aNKRs increases with the number of receptors per individual (**Figure 2**), because a large receptor number increases the chance of a host carrying very specific NKRs to have licensed receptors. This observation suggests that heterozygous hosts should have an advantage over homozygotes. We therefore hypothesized that heterozygous advantage must be selecting novel NKRs in our agent-based model, driving polymorphism of NKRs in the population.

To measure the polymorphism at population level, we use the Simpson's reciprocal index (SRI, see Materials and Methods). The SRI is a diversity measure that is equal to the total number of NKRs if they are equally distributed in the population, whereas the SRI is lower than that in a population where some alleles dominate (34).

The initial polymorphism of aNKRs (i.e., SRI = 10) increases over time (**Figure 4G**; black line), reflecting that a high number of aNKRs provides indeed an advantage. But surprisingly, this is not the case for iNKRs, where the diversity decreases to SRI = 7. Because the agent-based model considers a limited number of MHC molecules in the population, the specificity that the iNKRs evolve in the simulations is lower than that observed in the analytical model (i.e., $p_{i,\text{simulations}} \approx 0.18$ compared to $p_{i,\text{analytical}} = 0.10$) (see **Figure 2**). With this specificity value that is slightly lower than expected, all haplotypes tend to cover the entire (finite) MHC space, making it possible to have at least one licensed receptor. As a result, these populations can be well protected with a lower number of receptors. Therefore, there is little heterozygous advantage in populations having only iNKRs, resulting in a low level of polymorphism. Thus, the agent-based model finds a different solution for an optimal protection: it evolves contracted haplotypes (i.e., composed only of seven receptors) with slightly less specific iNKRs than expected.

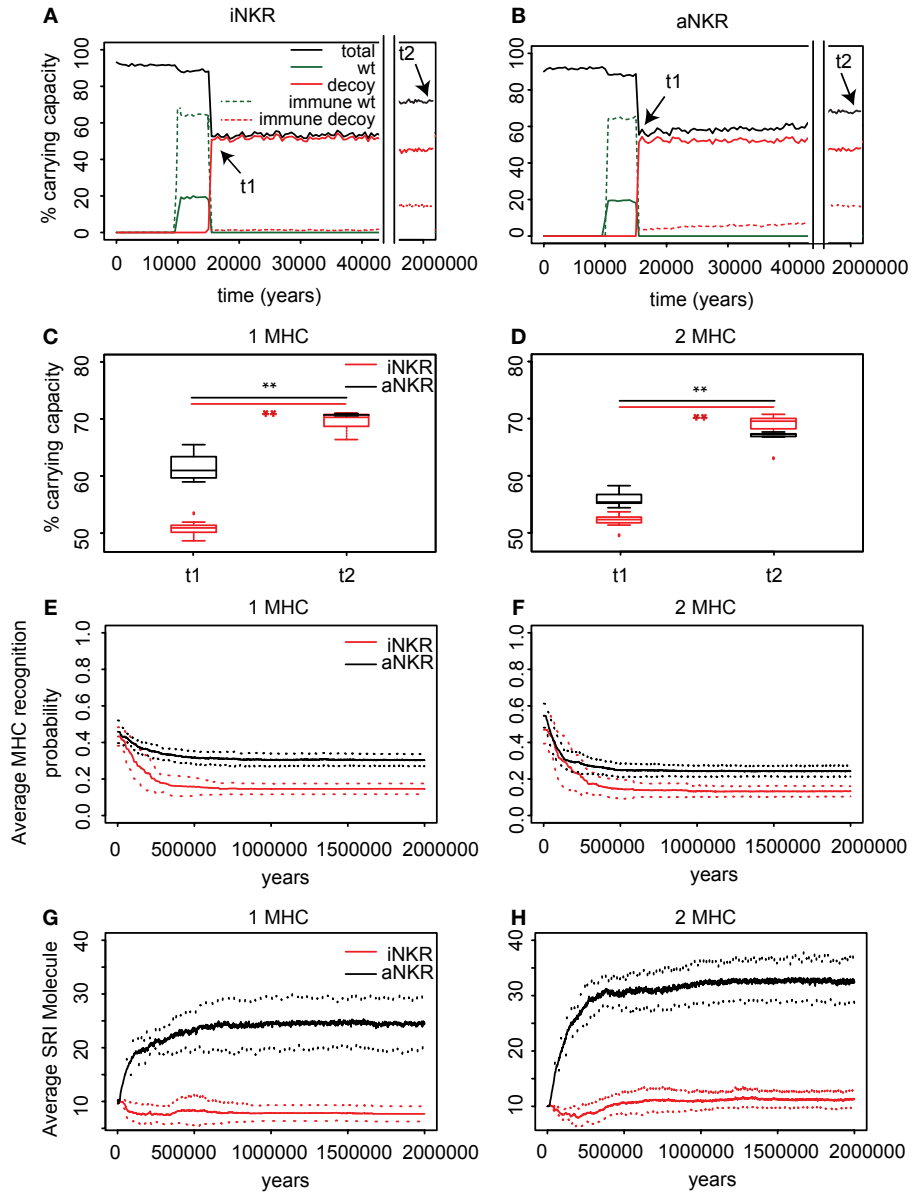


FIGURE 4 | Agent-based model confirms probabilistic model. **(A)** A host population having only iNKRs was inoculated with a wild-type virus (green lines) after a period of $t = 5000$ years (green solid lines show the chronically infected individuals and the dashed lines the immune individuals). Ten thousand years after the initial epidemic (i.e., t_1), we allowed for the evolution of decoy viruses (red lines). During the wild-type infection, most individuals recover (dashed green line). In contrast, almost none of the individuals are initially capable of clearing a CMV-like infection (red dashed line), resulting in a large decrease of the total population size (black line). **(B)** A host population having only aNKRs is initially better protected against decoy viruses, resulting in a higher fraction of the population clearing the infection, and a lower decrease of the total population size. **(A,B)** Show single representative simulations. **(C)** The average population size during the initial spread of decoy viruses (t_1) is lower than that at the end of the simulations (i.e., $t_2 = 3$ million years), indicating that over time, the populations learn to cope with the viral infection. Individuals in simulations considering only aNKRs (black) are initially better protected than those in simulations considering only iNKRs (red). In these simulations, all hosts carry only one MHC locus. **(D)** The initial advantage that aNKRs have over iNKRs receptors decreases in simulations considering two MHC loci per individual. **(E)** The probability of iNKRs recognizing any random MHC molecule in the population decreases over time (red line), indicating that more specific receptors are being selected for. In contrast, aNKRs (black line) do not evolve such high degree of specificity. **(F)** aNKRs evolve to become more specific in simulations where individuals have two MHC loci. **(G)** The degree of NKR polymorphism (expressed as the SRI score) increases in time, as a result of the evolved higher specificity. **(H)** SRI score in simulations considering two MHC loci. In **(C,D)**, the boxes represent the interquartile range, and the thick horizontal lines the median out of ten simulations (**represent p values < 0.005 and were calculated using the Mann–Whitney U test). In **(E–H)**, the solid lines represent the average out of ten simulations, and the dashed lines are the standard deviation.

Table 2 | Levels of protection against a decoy virus in the agent-based model.

	No. of $aNKR_{binding} = 0$	No. of $aNKR_{binding} > 0$
No. of $iNKR_{binding} = 0$	$p_{dec,1}$	$p_{dec,1}$
No. of $iNKR_{binding} > 0$	$p_{dec,0}$	$p_{dec,1}$

$aNKR_{binding}$ and $iNKR_{binding}$ refer to the number of iNKRs and aNKRs binding the decoy molecule, respectively. The receptors here are considered to be licensed.

PROTECTION DEPENDS ON THE NUMBER OF MHC LOCI

To confirm our results concerning the dependency on MHC loci number, we also perform simulations with individuals having two MHC loci. An increasing number of MHC loci has a large effect on the protection provided by aNKRs. Although these populations are initialized with intermediate specific NKRs, the initial protection is lower than in the population carrying only one MHC locus (**Figure 4D**). For better protection, a higher specificity is required, and the selection for more specific aNKRs is stronger in these simulations (**Figure 4F**). As a result of the higher specificity, a larger number of receptors per haplotype are necessary to become licensed and to recognize the foreign decoy molecules. Therefore, the advantage of heterozygotes over homozygotes is larger in these populations, resulting in a higher degree of polymorphism (**Figure 4H**).

The protection and evolution of iNKRs is less sensitive to the number of MHC loci per individual. Like in the simulations considering one MHC locus, we observe a recovery of the population as more specific receptors are evolving (**Figures 4D,F**). Because the total number of MHC alleles is larger in populations having two MHC loci, more iNKRs per haplotype are necessary to have at least one licensed receptor. Hence, the total SRI score is higher in these simulations, than in the case of single MHC locus (**Figures 4G,H**; red line).

BASIN OF ATTRACTION: MIXED HAPLOTYPES CONTAINING A MAJORITY OF aNKRs

Finally, we performed simulations of populations having both iNKRs and aNKRs, in which we allow for the evolution of the specificity and also the receptor type. The initial specificity values for both receptor types was intermediate (i.e., $p \approx 0.4$) and we initialized the genotypes with a random number of activating and inhibitory receptors.

After the appearance of decoy viruses, the populations suffered similar effects to those having only iNKRs and aNKRs. The population size decreases dramatically at first, and with time it recovers. The final population size is higher than in the simulations considering only one type of receptor, approaching 70% of the carrying capacity (**Figure 5A**) because mixed haplotypes protect better than only one type of receptors. At the end of the simulations, we observe more aNKRs than iNKRs per haplotype (**Figure 5B**), i.e., the final haplotypes are composed on average of 6 aNKRs and 4 iNKRs. In agreement with the predictions of the analytical model, both receptor types evolve high specificity (i.e., $p \leq 0.35$), and a high polymorphism (**Figures 5C,D**). Summarizing, the agent-based model confirms the prediction of the probabilistic model.

CONCLUSION AND DISCUSSION

Our mathematical model predicts the optimal protection level provided by inhibiting and activating NKRs against viruses expressing MHC-like molecules. Haplotypes composed only of iNKRs detect the viral infection within a small range, requiring high specificity and large haplotype size. In contrast, the maximal protection is reached for intermediate specificity values and at a smaller haplotype size in individuals having only aNKRs. Mixed haplotypes, i.e., haplotypes carrying both iNKRs and aNKRs offer the highest protection.

All these results are dependent on the number of MHC loci per individual. With increasing MHC loci, aNKRs lose their ability to become licensed and thus provide little or no protection. In contrast, haplotypes composed only of iNKR have a higher chance of having licensed receptors when the number of MHC loci is increased. In this case, the protection level is maximal already with a contracted NKR haplotype. Thus, there seem to be several combinations of MHC–NKR genotypes that provide maximal protection. A high protection is reached with a simple MHC complex and a high number of NKR genes. With increasing complexity of the MHC, a contracted NK complex is sufficient to render protection. These last results are particularly interesting, as they provide a possible explanation of the differences in KIR and MHC gene content across primate species (35), and the expansion of new KIR lineages corresponding to the contraction of the MHC gene complex, thus illustrating the co-evolution of MHC class I and KIRs (36).

The model described here is inspired by viruses evolving decoys. However, its main outcome, namely the requirement for specificity, might be more general than the defense against such decoy viruses. Studies have shown that viral infections can change the repertoire of peptides presented by MHC class I molecules (37), and that these different peptides affect the NKR–MHC interactions, perturbing the binding of iNKRs and leading to NK cell activation (38). In such cases, specific recognition of the changes in peptide repertoire by NK cells seems advantageous for the host. Also, the specificity ranges obtained in our model for mixed haplotypes (**Figure 3E**) are similar to those observed in reality, with iNKRs having a specificity of 0.2. This corresponds to the four mutually exclusive epitopes that have been detected so far for inhibitory KIRs in humans: HLA-A11, -Bw4, -C1, and -C2.

The exact role of aNKRs remains intriguing. Since only a few aNKRs tend to recognize MHC class I molecules (39), we speculate that aNKRs could specifically recognize new ligands expressed upon viral infection (e.g., decoy molecules or stress ligands). Our model predicts that to face the challenge of not recognizing self but specifically recognize foreign antigens, aNKRs do not need to be so specific. Indeed, the haplotype providing the highest protection is a combined haplotype composed of more aNKRs than iNKRs, which disagrees with the most primate KIR haplotypes (36). Most primate NKRs are inhibitory, and activating receptors have been linked to selection pressure induced by reproduction (36). Our model predicts that aNKRs should evolve to an intermediate specificity upon CMV-like infections. However, not many activating ligands have been identified yet, and it remains puzzling what other roles aNKRs might play.

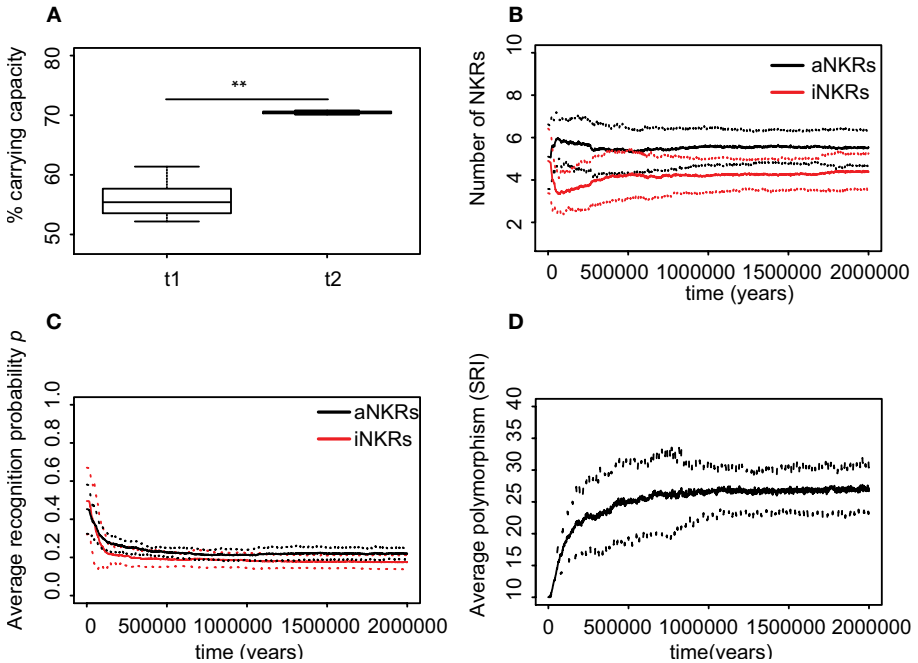


FIGURE 5 | Mixed haplotypes offer the highest protection. A host population having iNKRs and aNKRs was inoculated with a wild-type virus after a period of $t = 5000$ years; we allowed for the evolution of decoy viruses 10,000 years after the initial epidemic (i.e., t_1). **(A)** The population size during the initial spread of decoy viruses (t_1) is lower than that at the end of the simulations (i.e., $t_2 = 3$ million years), indicating that over time, the population recovers from the viral infection. **(B)** The initial haplotype is composed of five iNKRs and five aNKRs. The number of aNKRs and iNKRs per haplotype varies over time, resulting in a selection for haplotypes with a larger activating potential. **(C)** The probability of NKRs recognizing any random MHC molecule in the population, decreases over time, indicating that more specific receptors are being selected for. **(D)** The degree of NKR polymorphism (expressed as the SRI score) increases in time, as a result of the heterozygote advantage due to the evolved higher specificity. Averages are taken out of ten different simulations. In **(A)**, the boxes represent the interquartile range, and the thick horizontal lines the median (**represent p values <0.005, and were calculated using the Mann–Whitney U test). In **(B–D)**, the solid lines represent the average out of ten simulations, and the dashed lines are the standard deviation.

The expansion of NKR superfamilies, presumably in order to gain resistance against pathogens, illustrates the high evolutionary complexity of NKRs. We aimed to fully understand the effects of a single possible driving force of this evolutionary process, namely that of a viral encoded MHC-like molecule. Therefore, we focused on modeling only the evolution of NKRs in a simple model, which requires simplifying assumptions. For instance, we fixed MHC polymorphism despite the evidence of the co-evolution between MHC class I and KIRs (35, 36). Given their different evolutionary timescales, i.e., that MHC molecules are older than both Ly49 and KIRs, we chose to model the expansion and contraction of NKR systems within an already existing MHC diversity. Additionally, we assumed that decoy viruses down-regulate the expression of all MHC molecules in the host. Even though we do not expect selective MHC down-regulation to affect the evolution of aNKRs (since activating receptors cannot detect missing-self), the evolution of iNKRs might be affected because more licensed iNKRs will be necessary to recognize a virus that down-regulates only one of the host's MHC-I molecules. Note that if the licensed repertoire of iNKRs is larger, these receptors should be even more specific to avoid being "fooled" by the decoy molecule. The exact effect of selective MHC down-regulation on the specificity of iNKRs is an open interesting question, which we are currently working on.

Other simplifying assumptions were also necessary, such as considering a global NKR repertoire and ignoring the synergy between NKRs or the direct interaction between immune cells. Additionally, we ignored mutational operators that conserve similarity between pre- and post-mutation receptors (e.g., point mutations), as we only model mutations that significantly change receptor functionality. Including point mutations, did not affect the results qualitatively (results not shown), however a longer evolutionary time was necessary to approach the same solution of specificity. Overall, since our main results are of a qualitative nature, it seems unlikely that relaxing any of these assumptions would affect our main results. Note also that our agent-based model is inspired on humans and KIRs, with the advantage of having realistic parameters for processes like birth and death. However, the model can be generalized to other species, and qualitatively it represents a model of the evolution of the expansion of the NKR complex.

All our analytical results were consistent with the agent-based model and our analysis allowed us to quantify the protection against an infection for both receptor types. It confirmed our previous results that iNKRs should become specific enough (28). Our new approach has shed light into the possible contribution that each receptor type confers upon infection, and allowed us to conclude that mixed haplotypes render the best protection.

MATERIALS AND METHODS

AGENT-BASED MODEL

The agent-based model consists of two types of actors (hosts and viruses) and three events (birth, death, and infection). This model is virtually identical to the one published in Ref. (28). Briefly, we screen all hosts in a random order during each time step of 1 week, and confront them to one of the randomly chosen events. Hosts age over time and at the end of each time step, their age, infection state, and type of infection is updated. This cycle is repeated for two million years to simulate long term evolution. All model parameters are given in **Table 1**.

We model simplified diploid individuals, carrying gene complexes for NKR and MHC class I. For simplicity, we consider 15 MHC alleles per locus, resembling the most common HLA alleles in the European population (40). NKRs and their ligands are modeled with randomly generated bit strings as a simplified representation of amino acids (41). If the longest adjacent complementary match between two strings exceeds a threshold L, we allow for the receptor to interact with its ligand. Thus, the threshold L determines the specificity of each receptor: a receptor with a small L value will be very degenerate and the probability of a random NKR to recognize a random MHC molecule will be $p \approx 1$. In contrast, a receptor with a large L value will be specific, and accordingly, the probability of this receptor binding any MHC molecule in the population will be $p \sim 0$ [for a detailed description, see Ref. (28)].

RECEPTOR TYPES

In the present model, we allow for the evolution of aNKRs. When a novel NKR is generated, a random L value between 1 and 16 is assigned to it, and its type (i.e., whether it is activating or inhibitory) is also randomly chosen. Thus, each receptor has its particular specificity and functionality. Receptors with L values larger than 13 will usually not recognize any MHC molecules in the population, and are typically not functional. Genes encoding such non-functional NKRs are considered to be pseudogenes. Haplotypes containing pseudogenes are effectively shorter than haplotypes composed of fully functional NKRs. Thus, we can model the contraction and expansion of the NKR gene complex.

VIRAL INFECTIONS

In our simulations, we consider one wild-type virus and several "decoy viruses," i.e., viruses expressing MHC decoys. Each virus comes with a viral load, which is implemented as an increase of the host's death rate, VL_i depending on the infection state i (see **Table 1**), and a probability of clearing the infection p_{wt} and $p_{dec,s}$ for the wild-type and the decoy viruses, respectively. A decoy virus down-regulates the expression of all MHC molecules in that host, and encodes one MHC-like molecule. The evolution of decoy molecules is modeled by allowing the virus to adopt a randomly selected MHC molecule from its host with a rate μ_v. The virus keeps this decoy for the rest of the simulation. Because we fix the MHC polymorphism to 15 alleles per locus, the maximal number of decoy proteins that can evolve in the population is 15 for the simulations considering 1 MHC locus, and 30 for those considering two MHC loci.

We consider different levels of protection against a decoy virus, depending on the success of the virus to escape the NK cell response, s. If at least one of the licensed iNKR binds to the decoy molecule, there will be an inhibitory signal, the host will not be able to detect "missing-self," and the decoy virus will be successful. Similarly, if none of the licensed aNKRs recognizes the decoy molecule, the decoy virus will evade the NK cell response. Thus, none of the iNKRs or at least one aNKRs should bind the decoy molecule to render protection (**Table 2**). We model the immune escape by setting the probability of clearing the infection to zero, letting the host become chronically infected. In the case that a decoy is not successful, the host will be able to detect "missing-self." Since this virus is nevertheless able to evade the response from T cells (due to the MHC down-regulation), the probability of clearing the infection is lower than that of the wild-type virus ($p_{wt} = 0.85$). The resulting probability of clearing the infection is described by:

$$p_{dec,s} = \begin{cases} 0, & \text{if } s = 0 \ \left(\text{successful decoy}\right) \\ 0.5, & \text{if } s = 1 \ \left(\text{unsuccessful decoy}\right). \end{cases} \quad (5)$$

The rest of the parameters defining the infection dynamics and immune escape of the decoy viruses (i.e., time of infection, immunity time, and transmission probabilities) were set like in Ref. (28) and are described in **Table 1**.

MUTATION

During each birth event, NKRs undergo mutation with a probability, μ. To decrease computation time, we model mutation by randomly creating a new receptor with its particular specificity and signaling type. We do not consider other mutational operators, e.g., point mutations, recombination, deletion, or duplication.

We first perform simulations where only the specificity can evolve (i.e., a random value L is assigned to each new receptor), while the receptor type remains fixed. Hereby, we are able to compare what the basin of attraction for the specificity will be, if a population has only aNKRs or only iNKRs. We also simulate populations with mixed haplotypes, by allowing the receptor type to mutate.

NK CELL EDUCATION

During the birth event, an NK cell education process takes place. Like in our probabilistic model explained above, iNKRs which recognize *at least one* of the MHC molecules within one individual, and aNKRs that fail to recognize *all* of the MHC molecules within the host, are set to be licensed. In our model, only the licensed repertoire of NKRs will participate in an NK cell response (**Figure 1**).

MODEL INITIALIZATION

The model is initialized with a host population of 4500 hosts, with random ages between 1 and 70 years corresponding to a uniform age distribution. After approximately 10 host generations, this age distribution corresponds to more modern age distributions with the majority of individuals having an age between 15 and 60.

At the start of every simulation, a gene pool for MHC alleles is generated, the size of which depends on the number of MHC loci per individual. It consists of 15 alleles in simulations

considering one MHC locus per individual, and of 30 alleles in those simulations considering two MHC loci per individual. To create the initial genome of each individual, MHC genes were randomly drawn from the pool, while ten NKRs with intermediate specificity ($2 \leq L \leq 4$, i.e., $p \approx 0.4$) were generated. Thus, the initial haplotypes did not contain any pseudogenes. In the simulations considering mixed NKR haplotypes, the initial genes can be both activating and inhibitory. The type of each receptor was randomly chosen as explained above, resulting in approximately 50% of the receptors being activating. All individuals were initialized with the same NKR haplotype, but with different MHC genes.

GENETIC DIVERSITY

The Simpson's Index is a measurement of diversity that can be interpreted as the probability that two randomly chosen receptors from two random hosts in the population are identical (34). The reciprocal of the Simpson's Index defines a "weighted" diversity. The SRI was calculated as follows: $\text{SRI} = \frac{1}{\sum_{i=1}^{N} f_i^2}$, where f_i is the frequency of the receptor i over all NKRs in the population, and N is the total number of unique NKRs.

IMPLEMENTATION

The model was implemented in the C++ programing language. We considered populations with haplotypes composed of only aNKRs, only iNKRs, or both. In every scenario, we compared the effects of one or two MHC loci per individual. For each of these settings, we performed ten simulations for 2 million years. The code is available upon request.

ACKNOWLEDGMENTS

We thank Chris van Dorp, Leïla Perie, and Hanneke van Deutekom for helpful discussions and carefully reading the manuscript. We also thank Oussama Jarrousse and Johannes Textor for their technical support during the development of the code. This work was financially supported by the CLS program of the Netherlands Organization for Scientific Research (NWO), grant 635.100.025. This study was also supported by the "Virgo consortium," funded by the Dutch government (project number FES0908) and by the "Netherlands Genomics Initiative (NGI)" (project number 050-060-452). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Lanier LL. Evolutionary struggles between NK cells and viruses. *Nat Rev Immunol* (2008) **8**(4):259–68. doi:10.1038/nri2276
2. Lanier LL. NK cell recognition. *Annu Rev Immunol* (2005) **23**:225–74. doi:10.1146/annurev.immunol.23.021704.115526
3. Ljunggren HG, Kärre K. In search of the "missing self": MHC molecules and NK cell recognition. *Immunol Today* (1990) **11**(7):237–44. doi:10.1016/0167-5699(90)90097-S
4. O'Callaghan CA. Molecular basis of human natural killer cell recognition of HLA-E (human leukocyte antigen-E) and its relevance to clearance of pathogen-infected and tumour cells. *Clin Sci (Lond)* (2000) **99**(1):9–17. doi:10.1042/CS19990334
5. Braud VM, Allan DS, O'Callaghan CA, Söderström K, D'Andrea A, Ogg GS, et al. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* (1998) **391**(6669):795–9. doi:10.1038/35869
6. Shum BP, Flodin LR, Muir DG, Rajalingam R, Khakoo SI, Cleland S, et al. Conservation and variation in human and common chimpanzee CD94 and NKG2 genes. *J Immunol* (2002) **168**(1):240–52.
7. Trowsdale J, Barten R, Haude A, Stewart CA, Beck S, Wilson MJ. The genomic context of natural killer receptor extended gene families. *Immunol Rev* (2001) **181**:20–38. doi:10.1034/j.1600-065X.2001.1810102.x
8. Moesta AK, Parham P. Diverse functionality among human NK cell receptors for the C1 epitope of HLA-C: KIR2DS2, KIR2DL2, and KIR2DL3. *Front Immunol* (2012) **3**:336. doi:10.3389/fimmu.2012.00336
9. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res* (2012) **22**(10):1845–54. doi:10.1101/gr.137976.112
10. Vilches C, Parham P. KIR: diverse, rapidly evolving receptors of innate and adaptive immunity. *Annu Rev Immunol* (2002) **20**:217–51. doi:10.1146/annurev.immunol.20.092501.134942
11. Kelley J, Walter L, Trowsdale J. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet* (2005) **1**(2):e27. doi:10.1371/journal.pgen.0010027
12. Guethlein LA, Older Aguilar AM, Abi-Rached L, Parham P. Evolution of killer cell Ig-like receptor (KIR) genes: definition of an orangutan KIR haplotype reveals expansion of lineage III KIR associated with the emergence of MHC-C. *J Immunol* (2007) **179**(1):491–504.
13. Anfossi N, André P, Guia S, Falk CS, Roetynck S, Stewart CA, et al. Human NK cell education by inhibitory receptors for MHC class I. *Immunity* (2006) **25**(2):331–42. doi:10.1016/j.immuni.2006.06.013
14. Wilhelm BT, Gagnier L, Mager DL. Sequence analysis of the ly49 cluster in C57BL/6 mice: a rapidly evolving multigene family in the immune system. *Genomics* (2002) **80**(6):646–61. doi:10.1006/geno.2002.7004
15. Wilhelm BT, Mager DL. Rapid expansion of the Ly49 gene cluster in rat. *Genomics* (2004) **84**(1):218–21. doi:10.1016/j.ygeno.2004.01.010
16. Körner C, Altfeld M. Role of KIR3DS1 in human diseases. *Front Immunol* (2012) **3**:326. doi:10.3389/fimmu.2012.00326
17. Pelak K, Need AC, Fellay J, Shianna KV, Feng S, Urban TJ, et al. Copy number variation of KIR genes influences HIV-1 control. *PLoS Biol* (2011) **9**(11):e1001208. doi:10.1371/journal.pbio.1001208
18. Martin MP, Gao X, Lee J-H, Nelson GW, Detels R, Goedert JJ, et al. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet* (2002) **31**(4):429–34. doi:10.1038/ng934
19. Zhi-Ming L, Yu-lian J, Zhao-lei F, Chun-xiao W, Zhen-fang D, Bing-chang Z, et al. Polymorphisms of killer cell immunoglobulin-like receptor gene: possible association with susceptibility to or clearance of hepatitis B virus infection in Chinese Han population. *Croat Med J* (2007) **48**(6):800–6. doi:10.3325/cmj.2007.6.800
20. López-Vázquez A, Rodrigo L, Martínez-Borra J, Pérez R, Rodríguez M, Fdez-Morera JL, et al. Protective effect of the HLA-Bw4I80 epitope and the killer cell immunoglobulin-like receptor 3DS1 gene against the development of hepatocellular carcinoma in patients with hepatitis C virus infection. *J Infect Dis* (2005) **192**(1):162–5. doi:10.1086/430351
21. Besson C, Roetynck S, Williams F, Orsi L, Amiel C, Lependeven C, et al. Association of killer cell immunoglobulin-like receptor genes with Hodgkin's lymphoma in a familial study. *PLoS One* (2007) **2**(5):e406. doi:10.1371/journal.pone.0000406
22. Hiby SE, Apps R, Sharkey AM, Farrell LE, Gardner L, Mulder A, et al. Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. *J Clin Invest* (2010) **120**(11):4102–10. doi:10.1172/JCI43998
23. Smith HR, Heusel JW, Mehta IK, Kim S, Dorner BG, Naidenko OV, et al. Recognition of a virus-encoded ligand by a natural killer cell activation receptor. *Proc Natl Acad Sci U S A* (2002) **99**(13):8826–31. doi:10.1073/pnas.092258599
24. Arase H, Mocarski ES, Campbell AE, Hill AB, Lanier LL. Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors. *Science* (2002) **296**(5571):1323–6. doi:10.1126/science.1070884
25. Abi-Rached L, Parham P. Natural selection drives recurrent formation of activating killer cell immunoglobulin-like receptor and Ly49 from inhibitory homologues. *J Exp Med* (2005) **201**(8):1319–32. doi:10.1084/jem.20042558
26. Arase H, Lanier LL. Virus-driven evolution of natural killer cell receptors. *Microbes Infect* (2002) **4**(15):1505–12. doi:10.1016/S1286-4579(02)00033-3
27. Sun JC, Lanier LL. The natural selection of herpesviruses and virus-specific NK cell receptors. *Viruses* (2009) **1**(3):362. doi:10.3390/v1030362
28. Carrillo-Bustamante P, Kesmir C, de Boer RJ. Virus encoded MHC-like decoys diversify the inhibitory KIR repertoire. *PLoS Comput Biol* (2013) **9**(10):e1003264. doi:10.1371/journal.pcbi.1003264

29. Elliott JM, Yokoyama WM. Unifying concepts of MHC-dependent natural killer cell education. *Trends Immunol* (2011) **32**(8):364–72. doi:10.1016/j.it.2011.06.001
30. Chalifour A, Scarpellino L, Back J, Brodin P, Devèvre E, Gros F, et al. A role for cis interaction between the inhibitory Ly49A receptor and MHC class I for natural killer cell education. *Immunity* (2009) **30**(3):337–47. doi:10.1016/j.immuni.2008.12.019
31. Sun JC, Lanier LL. Tolerance of NK cells encountering their viral ligand during development. *J Exp Med* (2008) **205**(8):1819–28. doi:10.1084/jem.20072448
32. Fauriat C, Ivarsson MA, Ljunggren HG, Malmberg KJ, Michaëlsson J. Education of human natural killer cells by activating killer cell immunoglobulin-like receptors. *Blood* (2010) **115**(6):1166–74. doi:10.1182/blood-2009-09-245746
33. Carnes BA, Holden LR, Olshansky SJ, Witten MT, Siegel JS. Mortality partitions and their relevance to research on senescence. *Biogerontology* (2006) **7**(4):183–98. doi:10.1007/s10522-006-9020-3
34. Simpson E. Measurement of diversity. *Nature* (1949) **163**:688. doi:10.1038/163688a0
35. Sambrook JG, Bashirova A, Palmer S, Sims S, Trowsdale J, Abi-Rached L, et al. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res* (2005) **15**(1):25–35. doi:10.1101/gr.2381205
36. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol* (2013) **13**(2):133–44. doi:10.1038/nri3370
37. Hickman HD, Luis AD, Bardet W, Buchli R, Battson CL, Shearer MH, et al. Cutting edge: class I presentation of host peptides following HIV infection. *J Immunol* (2003) **171**(1):22–6.
38. Fadda L, O'Connor GM, Kumar S, Piechocka-Trocha A, Gardiner CM, Carrington M, et al. Common HIV-1 peptide variants mediate differential binding of KIR3DL1 to HLA-Bw4 molecules. *J Virol* (2011) **85**(12):5970–4. doi:10.1128/JVI.00412-11
39. Moesta AK, Graef T, Abi-Rached L, Older Aguilar AM, Guethlein LA, Parham P. Humans differ from other hominids in lacking an activating NK cell receptor that recognizes the C1 epitope of MHC class I. *J Immunol* (2010) **185**(7):4233–7. doi:10.4049/jimmunol.1001951
40. Meyer D, Singe R, Mack S, Lancaster A, Nelson M, Erlich H, et al. Single locus polymorphism of classical HLA genes. In: Hansen JA, editor. *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference.* (Vol. 1), Seattle, WA: IHWG Press (2007). p. 653–704.
41. Farmer JD, Packard NH, Perelson AS. The immune system, adaptation, and machine learning. *Physica D* (1986) **22**:187–204. doi:10.1016/0167-2789(86)90240-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 September 2013; accepted: 15 January 2014; published online: 30 January 2014.

*Citation: Carrillo-Bustamante P, Keşmir C and de Boer RJ (2014) Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus. Front. Immunol. **5**:20. doi: 10.3389/fimmu.2014.00020*

This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Carrillo-Bustamante, Keşmir and de Boer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

GLOBAL REPERTOIRE OF LICENSED NKRs

Instead of considering individual NK cells expressing a tuned set of NKRs, we model a global repertoire of "licensed" NKRs in each host. Hereby, we assume that the expression of at least one "licensed" receptor is sufficient to educate NK cells, and that those functional NK cells can become activated upon viral infection. This assumption needs further explanation.

If an aNKR recognizes some self-molecule (or self-MHC), all developing NK cells expressing that receptor should additionally express at least one iNKR recognizing self to prevent self-reactivity. As a consequence, all NK cells expressing that aNKR should never become activated, even if a virus encodes a ligand which engages that aNKR. Therefore, we call such an aNKR "unlicensed": it will never contribute to detect a viral infection. In contrast, if an aNKR that does not recognize any self-molecule, all developing NK cells expressing that receptor will not be influenced by it, and these cells will express other iNKRs and aNKRs that balance their self-reactivity. NK cells expressing that aNKR will become activated when a virus expressing its ligand engages that aNKR. Therefore we call such an aNKR "licensed." Any NK cell expressing this aNKR should breach its activation threshold, expand, and protect, when the ligand is present.

Now consider an iNKR that recognizes some self-MHC. All developing NK cells expressing that receptor will be tuned to balance its self-reactivity. As a consequence, all NK cells expressing that iNKR should become activated when a virus down-regulates this particular MHC. Therefore, we call such an iNKR "licensed": it detects MHC down-regulation. However, for an iNKR that does not recognize any self-MHC, the developing NK cells expressing that receptor will not be influenced by that iNKR. Consequently, these cells will express other iNKRs and other aNKRs in order to balance their self-reactivity. Therefore, NK cells expressing that iNKR will not become activated by this iNKR when a virus down-regulates self-MHC. Therefore, we call such an iNKR "unlicensed": it would never contribute to the detection of virus infected cells.

What happens with the NK cells expressing such an unlicensed iNKR if it happens to recognize something else on virus infected cells, e.g., a decoy? Then the iNKR should deliver an extra inhibitory signal to those cells. Consequently, these NK cells cannot expand and protect, even if their other iNKR detect MHC down-regulation, or if their other aNKR detect a new ligand. Hence, such an unlicensed iNKR will not protect, and can be considered to be non-functional at the repertoire level.

frontiers in
IMMUNOLOGY

GENERAL COMMENTARY
published: 05 January 2015
doi: 10.3389/fimmu.2014.00663

Corrigendum: Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus

Paola Carrillo-Bustamante*, Can Keşmir and Rob J. De Boer

Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, Netherlands
**Correspondence: p.carrillobustamante@uu.nl*

Edited by:
Ramit Mehr, Bar-Ilan University, Israel

Reviewed by:
Becca Asquith, Imperial College London, UK
Jayajit Das, The Ohio State University, USA

Keywords: agent-based modeling, NK cell receptors, evolution, CMV infection, models, theoretical

A corrigendum on

Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus
by Carrillo-Bustamante P, Keşmir C, de Boer RJ. Front Immunol (2014) 5:20. doi:10.3389/fimmu.2014.00020

We found a minor implementation error in the NK cell education process of our agent-based model. Fortunately, this hardly affected our main results and conclusions. The main difference lies in the polymorphism of iNKRs, as our new results show that it increases substantially over time similar to that of aNKRs. For reasons of accuracy and reproducibility, we here provide the corrected Figures and paragraphs (underlined).

POPULATIONS HAVING ONLY ANKRS EVOLVE A LARGER NKR POLYMORPHISM THAN POPULATIONS WITH ONLY INKRS

Our probabilistic model predicts that the protection by iNKRs and aNKRs increases with the number of receptors per individual (**Figure 2**), because a large receptor number increases the chance of a host carrying very specific NKRs to have licensed receptors. This observation suggests that heterozygous hosts should have an advantage over homozygotes. We therefore hypothesized that heterozygous advantage must be selecting novel NKRs in our agent-based model, driving polymorphism of NKRs in the population.

To measure the polymorphism at population level, we use the Simpson's reciprocal index (SRI, see Material and Methods). The SRI is a diversity measure that is equal to the total number of NKRs if they are equally distributed in the population, whereas the SRI is lower than that in a population where some alleles dominate (34).

The initial polymorphism of aNKRs (i.e., SRI = 10) increases over time (**Figure 4G** black line), reflecting that a high number of aNKRs provides indeed an advantage. Similarly, the SRI score of iNKRs increases over time. Because each evolved iNKRs recognizes on average one MHC molecule in the population, there is selection for haplotypes that do not overlap in the MHC molecules they recognize. Thus, the heterozygote advantage is large in these populations, driving the diversity of iNKRs.

PROTECTION DEPENDS ON THE NUMBER OF MHC LOCI

To confirm our results concerning the dependency on MHC loci number, we also perform simulations with individuals having two MHC loci. An increasing number of MHC loci has a large effect on the protection provided by aNKRs. Although these populations are initialized with intermediate specific NKRs, the initial protection is lower than in the population carrying only one MHC locus (**Figure 4D**). For better protection, a higher specificity is required, and the selection for more specific aNKRs is stronger in these simulations (**Figure 4F**). As a result of the higher specificity, a larger number of receptors per haplotype is necessary to become licensed and to recognize the foreign decoy molecules. Therefore, the advantage of heterozygotes over homozygotes is larger in these populations, resulting in a higher degree of polymorphism (**Figure 4H**).

The protection and evolution of iNKRs is also sensitive to the number of MHC loci per individual. Like in the simulations considering one MHC locus, we observe a recovery of the population as more specific receptors are evolving (**Figures 4D,F**). Hereby, the specificity evolved to even higher values, as the evolved iNKRs recognize on average <5% of all the MHC alleles in the population. Because of this high specificity and the larger number of MHC alleles in populations having two MHC loci, more iNKRs per haplotype are necessary to have at least one licensed receptor. Hence, the total SRI score is higher in these simulations, than in the case of single MHC locus (**Figures 4G,H** red line).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2014; accepted: 10 December 2014; published online: 05 January 2015.
Citation: Carrillo-Bustamante P, Keşmir C and De Boer RJ (2015) Corrigendum: Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus. Front. Immunol. 5:663. doi: 10.3389/fimmu.2014.00663
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2015 Carrillo-Bustamante, Keşmir and De Boer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

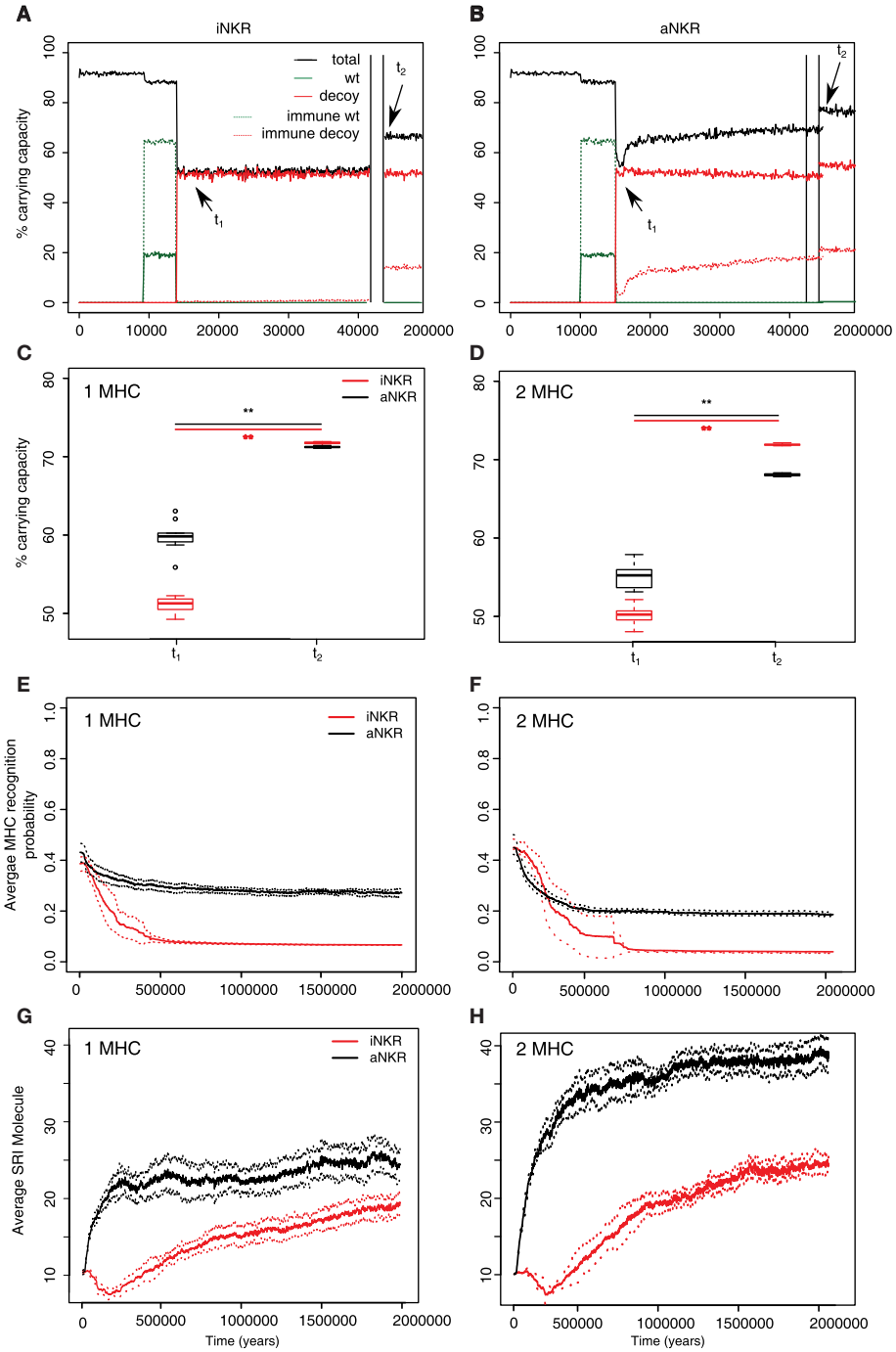


FIGURE 4 | Agent-based model confirms probabilistic model. **(A)** A host population having only iNKRs was inoculated with a wild type virus (green lines) after a period of $t = 5000$ years (green solid lines show the chronically infected individuals, and the dashed lines the immune individuals). 10,000 years after the initial epidemic (i.e., t_1), we allowed for the evolution of decoy viruses (red solid lines show the chronically infected individuals, and the dashed red lines the immune individuals). During the wild type infection, most individuals recover (dashed green line). In contrast, almost none of the individuals are initially capable of clearing a CMV-like infection (red dashed line), resulting in a large decrease of the total population size (black line). **(B)** A host population having only aNKRs is initially better protected against decoy viruses, resulting in a higher fraction of the population clearing the infection, and a lower decrease of the total population size. **(A,B)** show single representative simulations. **(C)** The average population size during the initial spread of decoy viruses (t_1) is lower than that at the end of the simulations (i.e., $t_2 = 3$ million years), indicating that over time, the populations learn to cope with the viral infection. Individuals in simulations considering only aNKRs (black) are initially better protected than those in simulations considering only iNKRs (red). In these simulations, all hosts carry only one MHC locus.

(Continued)

FIGURE 4 | Continued
(D) The initial advantage that aNKRs have over iNKRs receptors decreases in simulations considering two MHC loci per individual. **(E)** The probability of iNKRs recognizing any random MHC molecule in the population decreases over time (red line), indicating that more specific receptors are being selected for. In contrast, aNKRs (black line) do not evolve such high degree of specificity. **(F)** aNKRs evolve to become more specific in simulations where individuals have two MHC loci. **(G)** The degree of NKR polymorphism (expressed as the SRI score) increases in time, as a result of the evolved higher specificity. **(H)** SRI score in simulations considering two MHC loci. In **(C,D)**, the boxes represent the interquartile range, and the thick horizontal lines the median out of ten simulations (**represent p values <0.005, and were calculated using the Mann–Whitney U test). In **(E–H)**, the solid lines represent the average out of ten simulations, and the dashed lines are the SD.

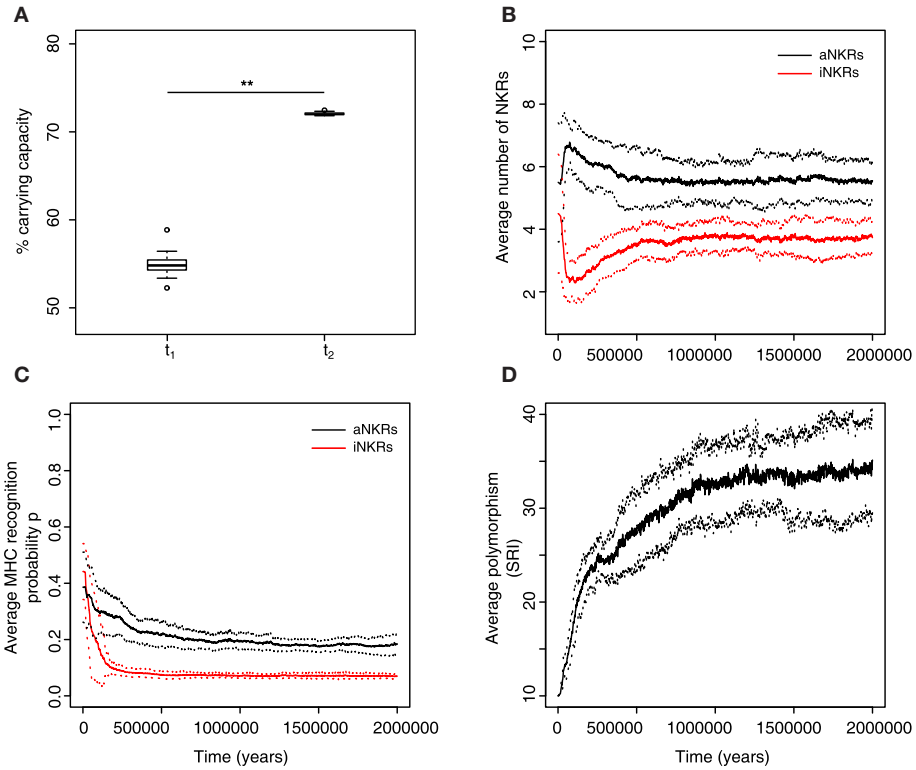


FIGURE 5 | Mixed haplotypes offer the highest protection. A host population having iNKRs and aNKRs was inoculated with a wild type virus after a period of $t = 5000$ years; we allowed for the evolution of decoy viruses 10,000 years after the initial epidemic (i.e., t_1). **(A)** The population size during the initial spread of decoy viruses (t_1) is lower than that at the end of the simulations (i.e., $t_1 = 3$ million years), indicating that over time, the population recovers from the viral infection. **(B)** The initial haplotype is composed of five iNKRs and five aNKRs. The number of aNKRs and iNKRs per haplotype varies over time, resulting in a selection for haplotypes with a larger activating potential. **(C)** The probability of NKRs recognizing any random MHC molecule in the population, decreases over time, indicating that more specific receptors are being selected for. **(D)** The degree of NKR polymorphism (expressed as the SRI score) increases in time, as a result of the heterozygote advantage due to the evolved higher specificity. Averages taken out of 10 different simulations. In **(A)**, the boxes represent the interquartile range, and the thick horizontal lines the median (**represent p values <0.005, and were calculated using the Mann–Whitney U test). In **(B–D)**, the solid lines represent the average out of ten simulations, and the dashed lines are the SD.

frontiers in IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 15 April 2014
doi: 10.3389/fimmu.2014.00172

An interaction library for the FcεRI signaling network

Lily A. Chylek[1,2], *David A. Holowka*[1], *Barbara A. Baird*[1]* and *William S. Hlavacek*[2]*

[1] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA
[2] Los Alamos National Laboratory, Theoretical Division, Center for Non-linear Studies, Los Alamos, NM, USA

Edited by:
Rob J. De Boer, Utrecht University, Netherlands

Reviewed by:
Christopher E. Rudd, University of Cambridge, UK
Grégoire Altan-Bonnet, Memorial Sloan-Kettering Cancer Center, USA

***Correspondence:**
Barbara A. Baird, Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA
e-mail: bab13@cornell.edu;
William S. Hlavacek, Los Alamos National Laboratory, Theoretical Division, Center for Nonlinear Studies, Los Alamos, NM 87545, USA
e-mail: wish@lanl.gov

Antigen receptors play a central role in adaptive immune responses. Although the molecular networks associated with these receptors have been extensively studied, we currently lack a systems-level understanding of how combinations of non-covalent interactions and post-translational modifications are regulated during signaling to impact cellular decision-making. To fill this knowledge gap, it will be necessary to formalize and piece together information about individual molecular mechanisms to form large-scale computational models of signaling networks. To this end, we have developed an interaction library for signaling by the high-affinity IgE receptor, FcεRI. The library consists of executable rules for protein–protein and protein–lipid interactions. This library extends earlier models for FcεRI signaling and introduces new interactions that have not previously been considered in a model. Thus, this interaction library is a toolkit with which existing models can be expanded and from which new models can be built. As an example, we present models of branching pathways from the adaptor protein Lat, which influence production of the phospholipid PIP_3 at the plasma membrane and the soluble second messenger IP_3. We find that inclusion of a positive feedback loop gives rise to a bistable switch, which may ensure robust responses to stimulation above a threshold level. In addition, the library is visualized to facilitate understanding of network circuitry and identification of network motifs.

Keywords: immunoreceptor signaling, IgE receptors (FcεRI), mast cells, knowledge engineering, computational modeling, network motifs, feed-forward loops, visualization

INTRODUCTION

Cell signaling plays a key part in regulation of the immune system. Adaptive immune responses are controlled by multichain immune recognition receptors, or immunoreceptors, which include the T cell receptor (TCR) (1), the B cell antigen receptor (BCR) (2), and the high-affinity receptor for IgE, which is also known as FcεRI (3). Each of these receptors is the gatekeeper of complex signaling machineries that translate extracellular stimuli into cellular responses. Individual interactions in immunoreceptor signaling systems have been studied extensively, and there is now a need to form a cohesive picture of how these interactions combine to mediate information processing. This need is driven in part by emerging data that reveal complex dynamical behaviors that arise from molecular interactions (4, 5), as well as by a growing appreciation of network features, such as crosstalk (6), which may only be apparent when one considers the interplay of multiple interactions.

Knowledge about signaling can be combined and synthesized into multiple forms, of which we employ two that are versatile and extensible: a visual map drawn in accordance with recommended standard conventions, and a rule-based model. The value of a standardized map, as opposed to an *ad hoc* cartoon, in depicting molecular interactions has been well appreciated: such maps can be used to organize information concisely, can be interpreted with minimal ambiguity, and can aid in logical analysis (7–11). After creation of a map, construction of a computational model can be viewed as the next level of information formalization (12). Through modeling, assumptions about molecular interactions (e.g., whether or not two interactions are competitive) are made more concrete and can thus be better assessed. In addition, modeling can extend our predictive capabilities when quantitative factors are important, enabling us to develop more sophisticated hypotheses. Modeling has become an increasingly important part of studies of immunoreceptor signaling (13).

Of the modeling frameworks that have been used to investigate biochemical systems, the framework of chemical kinetics is useful for studying dynamical behaviors that evolve on >1 ms time scales and that can be characterized using measurable parameters, such as protein copy numbers and binding rate constants. Among the modeling techniques of chemical kinetics is rule-based modeling (14), which provides a means to represent individual biomolecular sites, which is essential when, for example, different phosphorylation sites can recruit different binding partners (15). Rule-based modeling also enables simulation of the behavior of a large number of distinct chemical species. Myriad multicomponent protein complexes and protein phosphoforms, for example, can potentially arise in cell signaling systems and this complexity poses a challenge for other modeling techniques (16, 17). Rule-based models are built from executable rules. Rules in a model have a certain degree of interdependence, but tend to be more modular than the component parts used in other modeling techniques, such as ordinary differential equations (17). Thus, it is not only possible to formulate rules for a specific model, but to construct general rule libraries from which different models may be built.

To further our systems-level understanding of immunoreceptor signaling, we have developed a map and a rule library for

early signaling mediated by FcεRI, which shares features with other related immunoreceptors. The FcεRI signaling system has a special feature of experimental tractability because the receptor can be stimulated using structurally defined antigens (18–20), making it a valuable model system for understanding how signaling is initiated. Furthermore, FcεRI has been the subject of several past modeling studies that have elucidated early events following receptor crosslinking (21, 22), the flow of information during signaling (23), aggregation of adaptor proteins (24, 25), and the impact of ligand dose and binding kinetics on kinase activation (26, 27). Aspects of the models used in these studies form a foundation for the rule library presented here. The library extends previous work by adding rules for interactions not previously included in models for FcεRI signaling. Thus, the library serves as a bridge between past studies of relatively small scope, and potential future studies that integrate information about more network elements to, for example, analyze multiplexed signaling data (28). As a first example of library use, we present simulations of recruitment of signaling proteins to the adaptor Lat, which is phosphorylated in response to FcεRI stimulation (29).

METHODS

We developed a library of rules based on known protein–protein and protein–lipid interactions, which were identified through a survey of the FcεRI literature. The rules can be assembled into different sets to form different models that capture the chemical kinetics of FcεRI signaling with site-specific resolution (14, 16, 30). Here, the term "site" is used to refer to a generic functional site in a biomolecule, which in the case of a protein may be a domain, linear motif, or amino acid residue subject to post-translational modification. In a rule-based model, rules capture knowledge about biomolecular interactions of interest. The rules in a model specify what interactions can occur in a system and under what conditions these interactions occur. A rule provides necessary and sufficient conditions for testing its applicability, a definition of the consequences of an interaction, and a rate law. A detailed example of a rule is illustrated graphically in Figure S1 in Supplementary Material. Rules, in combination with parameters and initial conditions, can be processed to simulate the time-dependent behavior of a signaling system, including the time-dependent formation of protein complexes and post-translational modifications of proteins at specific sites. A benefit of a rule-based approach is that it enables concise specification and efficient simulation of models that include multivalent interactions and multi-site phosphorylation, which are two inherent characteristics of immunoreceptor signaling systems that are otherwise difficult or impossible to fully capture in a physicochemical model. We specified our library using a domain-specific language for rule-based modeling, the BioNetGen language (BNGL) (30), which is compatible with several software tools for simulation and analysis.

Our simulations are based on the law of mass action and an assumption of well-mixed reaction compartments. In the example model, the following compartments are considered implicitly: the cytosol, the plasma membrane, and the extracellular fluid surrounding a single-cell. Simulations were performed using CVODE (31), the built-in deterministic simulator of BioNetGen, which takes as input the ODEs derived from a rule-specified reaction network. Our illustrations of rules are based on published guidelines for model visualization (10) and were drawn with the help of a template available online (http://bionetgen.org/index.php/Extended_Contact_Maps).

In our bifurcation analyses, we found stable steady states through simulations that were started from arbitrary initial conditions or nearby steady states. The bifurcation parameter was an input signal taken in the model of interest to control the rate of activation of Syk and Fyn, which were each deactivated through a first-order process. Thus, as the input signal increases, so too do the steady-state levels of active Syk and Fyn. In simulations performed to find stable steady states, the bifurcation parameter was systematically varied from a low to high value, and vice versa.

To characterize signaling dynamics for specific observables (i.e., model outputs), we calculated rise time as the time required for the observable to reach 95% of its final steady-state value. For comparison between two models, a ratio of rise times was calculated.

RESULTS AND DISCUSSION

LIBRARY

In this section, we present a collection of rules, which can be viewed as a single model or as an assemblage of multiple models. Our main purpose is not to simulate the full set of interactions represented by these rules, but to formalize available knowledge about the FcεRI system to facilitate future modeling studies aimed at addressing specific questions. Rules in the library are provided in File S1 in Supplementary Material.

Rule-based models are compositional, meaning that rules can be specified somewhat independently of each other, enabling construction of new models from components of existing models. We have taken advantage of this feature to build on three previously reported models: one for ligand–receptor interactions and two for intracellular signaling. Below, we briefly review these models and the processes that they capture. A visual overview of the intracellular processes captured in the library is provided in **Figure 1**.

Initiation of signaling by FcεRI requires aggregation of receptors, which can be induced by reagents such as haptenated proteins and polymers, as well as by anti-receptor antibodies (32). Several models have been developed to investigate the interactions that lead to receptor aggregation. The model that we consider here is that of Xu et al. (33) for interactions of IgE-FcεRI with DNP–BSA, a multivalent antigen (haptenated protein). We chose this model because DNP–BSA is commonly used for stimulation of mast cells sensitized with anti-DNP IgE, and receptor aggregation induced by this antigen has been studied in detail (34). In this model, the effective valence of the ligand was taken to be two. The model includes transient hapten exposure, initial binding of a ligand to a receptor, crosslinking of neighboring receptors, and dissociation of ligand–receptor bonds. In this model, it was assumed that receptor sites (antigen-combining sites in cell-surface IgE) are equivalent and that the single-site dissociation rate constant is the same for both ligand sites, regardless of whether the second site is bound or free. Cyclic aggregates are not considered. For use in this study, the model was translated from its original form to rules, which was also done in another recent study (35). The model of Xu et al. is illustrated in **Figure 2**.

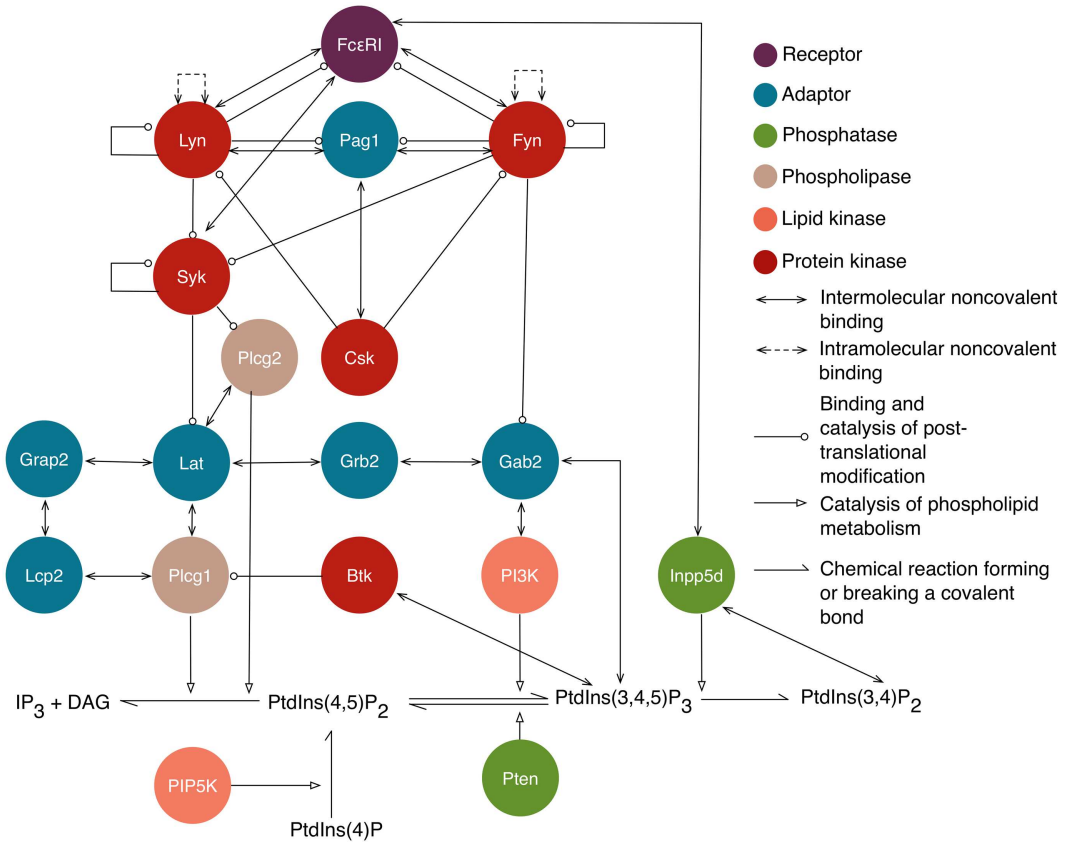


FIGURE 1 | An overview of intracellular signaling interactions included in the model/library for FcεRI signaling. Rules are included in the library for the interactions depicted here. Proteins are represented as circles that are color-coded according to their function, as indicated in the legend. Standard UniProt names are used, and we note that Grap2 is commonly known as Gads, Lcp2 is commonly known as Slp76, and Inpp5d is commonly known as Ship1. The legend also indicates the arrows that are used to represent different types of interactions and influences. Reactions of lipid species are illustrated at the bottom. Arrows from proteins that point to lipid reactions indicate that the reaction is catalyzed by that protein. Arrows from protein to lipid species indicate that the protein binds that lipid. Not shown are implicit phosphatase reactions that cause dephosphorylation of all sites that can be phosphorylated. Ligand–receptor interactions are shown in **Figure 2**. A subset of interactions is illustrated with site-specific detail in **Figure 3**.

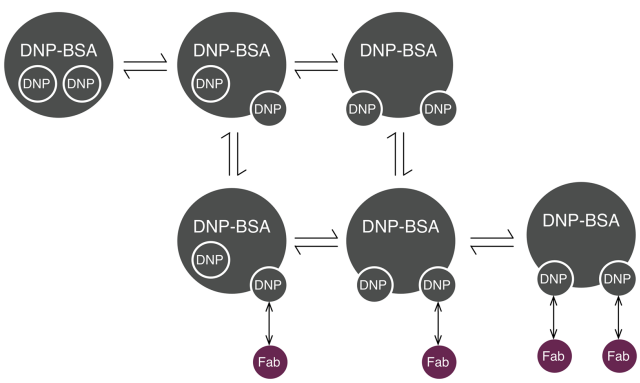


FIGURE 2 | Reaction scheme for DNP–BSA interactions with cell-surface IgE. BSA (bovine serum albumin) is haptenated with multiple DNP groups, which are assumed to transition between two states: inaccessible (represented as being inside the molecule) and accessible (represented as being on the edge of the molecule). Accessible DNP can bind Fab arms of IgE. Each IgE antibody has two Fab arms, and is thus bivalent.

Receptor aggregation initiates signaling by bringing receptors into proximity with the Src-family kinase (SFK) Lyn. Lyn's association with receptors may be facilitated by several complementary mechanisms, including regulation by the membrane lipid environment (36) and constitutive direct binding to FcεRI via Lyn's unique N-terminal domain (37). For simplicity, we explicitly model the latter mechanism because it allows the plasma membrane to be treated as well-mixed and has been formalized in past modeling studies (21, 22). Lyn mediates phosphorylation of other receptors in an aggregate, thereby generating binding sites for the SH2 domain of Lyn. In this model, FcεRI constitutively associates with the unique N-terminal domain of Lyn. Crosslinking of receptors enables Lyn to *trans* phosphorylate a second receptor at sites in the receptor's cytoplasmic subunits. These subunits, a β chain (Ms4a2) and a homodimer of two γ chains (Fcer1g), each contain an immunoreceptor tyrosine-based activation motif (ITAM). Each ITAM contains two (canonical) tyrosine residues that can be phosphorylated. The β chain contains an additional, non-canonical tyrosine in the middle of the ITAM sequence. In the original model, the tyrosines in the β chain were treated as a

single-site, as were tyrosines in the γ chains. Here, we consider the β chain's N-terminal (canonical) and middle (non-canonical) tyrosines separately because they are capable of recruiting distinct binding partners. The phosphorylated N-terminal tyrosine recruits Lyn to aggregated receptors via SH2 domain binding, and enhances Lyn's catalytic activity by disruption of an inhibitory intramolecular bond, forming a positive feedback loop. The non-canonical phosphotyrosine binds the lipid phosphatase Inpp5d (Ship1), which we will discuss below. The dually phosphorylated γ ITAM binds the tandem SH2 domains of the kinase Syk. Tyrosine residues in the linker region of Syk are phosphorylated by Lyn. Syk *trans* phosphorylates the activation loop in a second Syk molecule that is co-localized by being bound to cross-linked receptor, which constitutes positive feedback.

Rules for additional interactions among signaling proteins, which include mediators of negative regulation, were adapted from a model for BCR signaling (38). Lyn and a second SFK, Fyn, bind the transmembrane adaptor protein Pag1. Pag1 can then be phosphorylated by these kinases, generating additional binding sites for Lyn and Fyn, as well as for the kinase Csk. When co-localized on Pag1, Csk can phosphorylate Lyn and Fyn at an inhibitory C-terminal tyrosine. In this model, it was assumed that phosphorylation occurs in *cis*, meaning that Csk mediates phosphorylation of an SFK only when both are bound to the same Pag1 molecule. The C-terminal phosphotyrosine of an SFK forms an intramolecular bond with the SFK's SH2 domain, resulting in autoinhibition of the SFK's kinase domain.

The new rules of our library join the proximal signaling events described above to downstream processes that have not previously been considered in mechanistic models of FcεRI signaling. New rules are discussed in the sections that follow and are illustrated in **Figure 3**. The nomenclature and residue numbers used are consistent with UniProt conventions for rat proteins (39), because rat cells are commonly used in experimental studies of FcεRI signaling. If we view the rules of our library as constituting a single model, then the terminal output of the model is production of IP_3, which is a second messenger. Binding of IP_3 to its receptor on the endoplasmic reticulum leads to release of Ca^{2+} ions from intracellular stores, which is a key step for several processes in mast cell function, including degranulation and chemotaxis (40). Finally, we note that the interactions included in this library are not all unique to FcεRI signaling and are shared by pathways operative in TCR and BCR signaling. Thus, to facilitate identification of rules applicable to multiple pathways/cell types, in Table S1

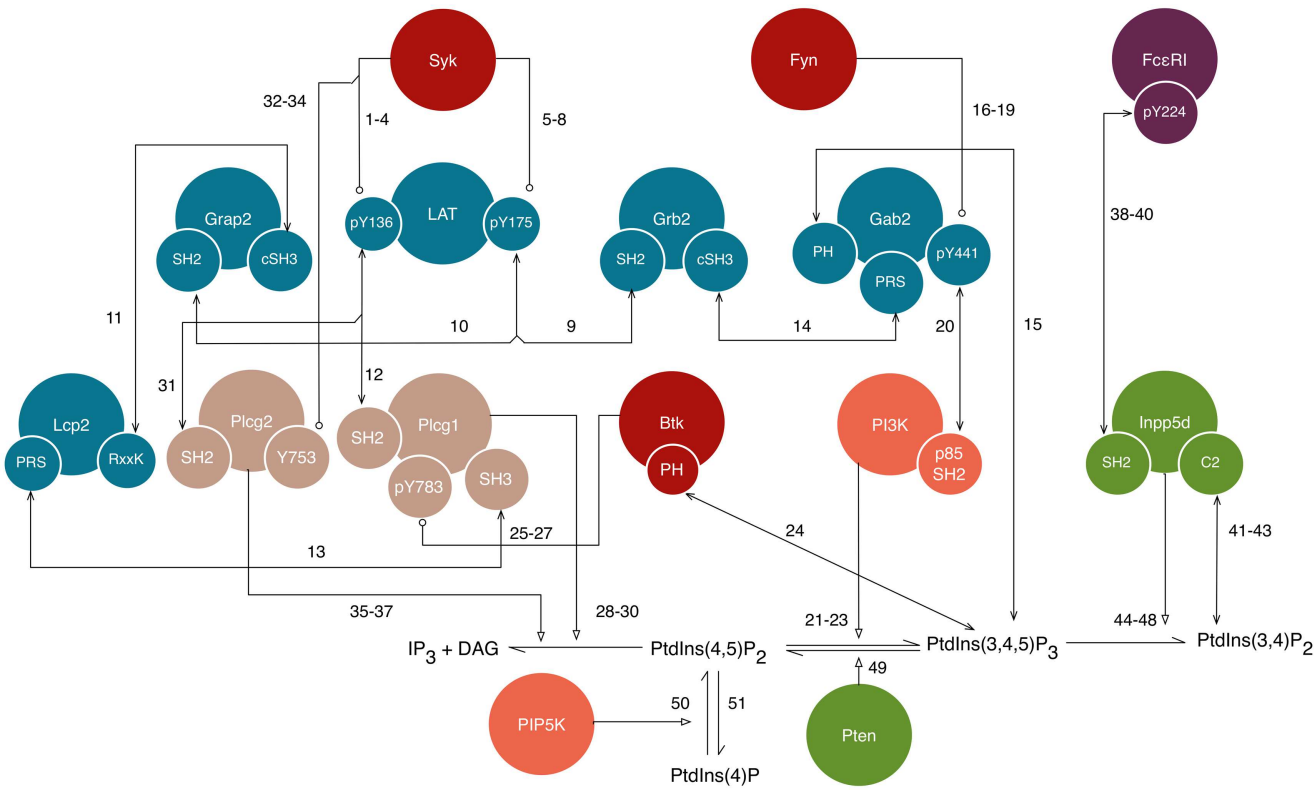


FIGURE 3 | A detailed illustration of new interactions included in the model for FcεRI signaling. This diagram shows a subset of the interactions shown in **Figure 1**, but with illustration of additional details, namely the sites responsible for interactions. Conventions for color-coding and arrow symbols are the same as in **Figure 1**. Large circles represent proteins. Small circles, overlapping the edges of large circles, represent protein sites/components, such as domains, motifs, and amino acid residues. Standard UniProt names are used, and we note that Grap2 is commonly known as Gads, Lcp2 is commonly known as Slp76, and Inpp5d is commonly known as Ship1. Arrows represent interactions and are numbered to correspond to the numbering of rules given in the text. Phosphatase activity is considered implicitly as dephosphorylation reactions that apply to all sites, and is not illustrated in this figure.

in Supplementary Material, we list protein–protein interactions included in the FcεRI library and whether each interaction is part of TCR and BCR signaling according to the NetPath database (41).

PHOSPHORYLATION OF LAT

Lat is a transmembrane, palmitoylated adaptor protein (42) that is involved in many signaling processes in both T cells and mast cells (43, 44). Syk phosphorylates Lat at multiple tyrosine residues (43), of which we focus on two: Y136 and Y175, which are better known as Y132 and Y191 in human Lat. Recent imaging studies suggest that Lat and the receptor become co-clustered after antigen-mediated receptor aggregation (45, 46). However, it is not clear if Syk-mediated phosphorylation of Lat takes place within the context of a signaling complex that co-localizes Syk and Lat, or if instead, Syk-mediated phosphorylation of Lat takes place through random collisions between Syk's kinase domain and tyrosine substrates in Lat that generate short-lived enzyme–substrate complexes, as in a Michaelis–Menten mechanism. It has previously been assumed that the latter mechanism holds (47) and we follow this approach, using rules capturing enzyme–substrate binding, dissociation, and catalysis. For example, the rules listed below, which are written using the conventions of BNGL (30), represent Syk-catalyzed phosphorylation of Y136 in Lat. Mass action kinetics are assumed. Bond indices are prefixed with the "!" symbol and internal state labels are prefixed with the "~" symbol. Here, internal state labels indicate whether a tyrosine residue is phosphorylated ("P") or unphosphorylated ("0").

```
(1) Syk(tSH2!+,PTK)+Lat(Y136~0) ->
    Syk(tSH2!+,PTK!1).Lat(Y136~0!1)
    kfSykLat
(2) Syk(PTK!1).Lat(Y136~0!1) -> Syk(PTK)+
    Lat(Y136~0) krSykLat
(3) Syk(PTK!1,Y519_Y520~P).Lat(Y136~0!1) ->
    Syk(PTK,Y519_Y520~P)+Lat(Y136~P)
    kpSykLat136_1
(4) Syk(PTK!1,Y519_Y520~0).Lat(Y136~0!1) ->
    Syk(PTK,Y519_Y520~0)+Lat(Y136~P)
    kpSykLat136_2
```

The first rule represents binding of Syk to Lat. In general, for rules in our library, protein components' names are consistent with terminology used in the biological literature. Here, the PTK component of Syk represents the protein tyrosine kinase domain of the protein. We assume that the interaction represented by Rule 1 only occurs when Syk is recruited to the plasma membrane, through binding of its tandem SH2 domains (tSH2) to phosphorylated FcεRI. Thus, the rule specifies that the tSH2 component must be bound for the reaction to occur (indicated by "!+"). The second rule represents the reverse reaction, which occurs independently of the binding state of Syk. Thus, the tSH2 component of Syk is not included in this rule. Rules 3 and 4 represent phosphorylation of Lat Y136 by Syk. These two rules differ in whether Syk is phosphorylated at its activation loop tyrosine residues Y519 and Y520, which are treated as a single site for simplicity. Phosphorylation of the activation loop enhances the catalytic activity of Syk (48). Rate constants consistent with this regulatory mechanism are given after each rule, and are assigned values in the "parameters" block of the model specification (File S1 in Supplementary Material). A similar set of rules are used to capture phosphorylation of Y175 in Lat.

```
(5) Syk(tSH2!+,PTK)+Lat(Y175~0) ->
    Syk(tSH2!+,PTK!1).Lat(Y175~0!1)
    kfSykLat
(6) Syk(PTK!1).Lat(Y175~0!1) -> Syk(PTK)+
    Lat(Y175~0) krSykLat
(7) Syk(PTK!1,Y519_Y520~P).Lat(Y175~0!1) ->
    Syk(PTK,Y519_Y520~P)+Lat(Y175~P)
    kpSykLat175_2
(8) Syk(PTK!1,Y519_Y520~0).Lat(Y175~0!1) ->
    Syk(PTK,Y519_Y520~0)+Lat(Y175~P)
    kpSykLat175_1
```

INTERACTIONS AMONG LAT AND ITS BINDING PARTNERS

Phosphorylated Y136 and Y175 have preferences for distinct binding partners, although crosstalk occurs between the pathways that branch from each site. Phosphorylated Y175 binds Grb2 and Grap2 (commonly known as Gads) (49), which are two related cytosolic adaptor proteins that each contain an SH2 domain flanked by two SH3 domains (50). These adaptors are also able to bind other sites in Lat, with Grb2 being more promiscuous (51), but for simplicity we focus on Y175. The interactions of Lat pY175 with Grb2 and Grap2, which are taken to be mutually exclusive, are modeled as follows:

```
(9) Lat(Y175~P)+Grb2(SH2) <->
    Lat(Y175~P!1).Grb2(SH2!1) kfLatGrb2,
    krLatGrb2
(10) Lat(Y175~P)+Grap2(SH2) <->
    Lat(Y175~P!1).Grap2(SH2!1) kfLatGrap2,
    krLatGrap2
```

These rules are nearly as general as possible, in that minimal molecular context is included on the left-hand side of either of these rules (i.e., the only requirements for a bond to form is availability of the cognate binding sites in each molecule). For this reason, a large number of distinct reactions are implicitly defined by each rule. This feature is a generic aspect of rules and what allows for concise model specification.

Grap2 binds Lcp2, which is also known as Slp76. This high-affinity interaction occurs through the SH3 domain of Grap2 and an unconventional RxxK motif in Lcp2 (52).

```
(11) Grap2(SH3)+Lcp2(RxxK) <->
    Grap2(SH3!1).Lcp2(RxxK!1) kfGrap2Lcp,
    krGrap2Lcp
```

Phosphorylated Y136 in Lat binds phospholipase Cγ1 (Plcg1) with high specificity (49), and the interaction is modeled with the following rule:

```
(12) Lat(Y136~P)+Plcg1(SH2) <->
    Lat(Y136~P!1).Plcg1(SH2!1) kfLatPlcg,
    krLatPlcg
```

Both of the tandem SH2 domains of Plcg1 contribute to co-localization of this enzyme with FcεRI upon stimulation (53), and there is evidence both SH2 domains are capable of binding Lat (54). However, for simplicity, we only consider a single SH2 domain in this model.

Plcg1 also interacts with Lcp2, via the SH3 domain of Plcg1 (55).

```
(13) Lcp2(PRS) + Plcg1(SH3) <-> Lcp2(PRS!1).
     Plcg1(SH3!1) kfLcp2Plcg1,krLcp2Plcg1
```

The final adaptor protein that we consider is Gab2. A linear motif in Gab2 can bind to the C-terminal SH3 domain of Grb2. We designate this motif as a proline-rich sequence (PRS), although its sequence differs from conventional SH3 binding motifs (56).

```
(14) Grb2(cSH3) + Gab2(PRS) <-> Grb2(cSH3!1).
     Gab2(PRS!1) kfGrb2Gab2,krGrb2Gab2
```

In addition, Gab2 can be recruited by binding of its PH domain to phosphatidylinositol 3,4,5-trisphosphate [PtdIns(3,4,5)P_3], also abbreviated as PIP_3, in the plasma membrane (57).

```
(15) PI345P3(headgroup) + Gab2(PH) <->
     PI345P3(headgroup!1).Gab2(PH!1)
     kfGab2Pip3,krGab2Pip3
```

The "headgroup" component in these rules represents the headgroup of the lipid, which is responsible for interactions with proteins.

RECRUITMENT OF PI3K TO Gab2

PI3K association with Gab2 is dependent on Gab2 phosphorylation. Gab2 is phosphorylated by Fyn (58), which we assume catalyzes phosphorylation through a Michaelis–Menten mechanism.

```
(16) Fyn(U!+,SH2,PTK) + Lat(Y175~P!1).
     Grb2(SH2!1,cSH3!2).Gab2(PRS!2,Y441~0) ->
     Fyn(U!+,SH2,PTK!3).Lat(Y175~P!1).
     Grb2(SH2!1,cSH3!2).Gab2(PRS!2,Y441~0!3)
     kfFynGab2
(17) Rec(b_Y210~P!4).Fyn(U,SH2!4,PTK) +
     Lat(Y175~P!1).Grb2(SH2!1,cSH3!2).
     Gab2(PRS!2,Y441~0) -> Rec(b_Y210~P!4).
     Fyn(U,SH2!4,PTK!3).Lat(Y175~P!1).
     Grb2(SH2!1,cSH3!2).Gab2(PRS!2,Y441~0!3)
     kfFynGab2
(18) Fyn(PTK!1).Gab2(Y441~0!1) ->
     Fyn(PTK) + Gab2(Y441~0) krFynGab2
(19) Fyn(PTK!1).Gab2(Y441~0!1) ->
     Fyn(PTK) + Gab2(Y441~P) kpFynGab2
```

The first two rules differ with respect to the mechanism by which Fyn is bound to a receptor. In the first rule, Fyn is taken to be bound by its unique domain (U). In the second rule, Fyn is taken to be bound by its SH2 domain.

Phosphorylated Gab2 binds the SH2 domain in the p85 subunit of PI3K (p85_SH2). Y441 of Gab2 lies in a consensus sequence for p85 binding (59).

```
(20) Gab2(Y441~P) + Pi3k(p85_SH2) <->
     Gab2(Y441~P!1).Pi3k(p85_SH2!1)
     kfGab2Pi3k,krGab2Pi3k
```

PI3K ACTIVITY

Once recruited, PI3K phosphorylates the 3rd position in the inositol ring of phosphatidylinositol 4,5-bisphosphate [PtdIns(4,5)P_2], also abbreviated as PIP_2, generating PIP_3.

```
(21) Lat(Y175~P!1).Grb2(SH2!1,cSH3!2).
     Gab2(PRS!2,Y441~P!3).Pi3k(p85_SH2!3,
     PI3Kc) + PI45P2(headgroup) ->
     Lat(Y175~P!1).Grb2(SH2!1,cSH3!2).
     Gab2(PRS!2,Y441~P!3).Pi3k(p85_SH2!3,
     PI3Kc!4).PI45P2(headgroup!4) kfPi3kPip2
(22) Pi3k(PI3Kc!1).PI45P2(headgroup!1) ->
     Pi3k(PI3Kc) + PI45P2(headgroup)
     krPi3kPip2
(23) Pi3k(PI3Kc!1).PI45P2(headgroup!1) ->
     Pi3k(PI3Kc) + PI345P3(headgroup)
     kpPi3k DeleteMolecules
```

In these rules, lipid phosphorylation is treated as consumption and production of different lipid species. For this reason, the BNGL keyword "DeleteMolecules" is used to indicate removal of reactant molecules (30).

Btk-MEDIATED ACTIVATION OF Plcg1

PtdIns(3,4,5)P_3 is a binding partner for multiple proteins, including the Tec-family kinase Btk, which is involved in activating Plcg1. The PH domain of Btk mediates this interaction.

```
(24) Btk(PH) + PI345P3(headgroup) <->
     Btk(PH!1).PI345P3(headgroup!1)
     kfBtkPip3,krBtkPip3
```

Recruited Btk can phosphorylate Plcg1 at sites that are associated with enhancement of phospholipase activity (60). In this way, pathways that branch from the two Lat phosphosites, Y136 and Y175, converge in contributing to IP_3 production.

```
(25) Btk(PH!+,PTK) + Plcg1(SH2!+,Y783~0) ->
     Btk(PH!+,PTK!1).Plcg1(SH2!+,Y783~0!1)
     kfBtkPlcg
(26) Btk(PTK!1).Plcg1(Y783~0!1) -> Btk(PTK) +
     Plcg1(Y783~0) krBtkPlcg
(27) Btk(PTK!1).Plcg1(Y783~0!1) -> Btk(PTK) +
     Plcg1(Y783~P) kpBtkPlcg
```

Plcg1 ACTIVITY

Plcg1 cleaves PtdIns(4,5)P_2 to generate the second messengers diacyl glycerol (DAG) and inositol 1,4,5-trisphosphate (IP_3) (61). The

cleavage reaction is taken to occur through a Michaelis–Menten mechanism:

```
(28) Plcg1(SH2!+,PLC)+PI45P2(headgroup) ->
     Plcg1(SH2!+,PLC!1).PI45P2(headgroup!1)
     kfPlcgPip2
(29) Plcg1(PLC!1).PI45P2(headgroup!1) ->
     Plcg1(PLC)+PI45P2(headgroup) krPlcgPip2
(30) Plcg1(PLC!1,Y783~P).PI45P2(headgroup!1)
     -> Plcg1(PLC,Y783~P)+IP3()+DAG()
     kcPlcg DeleteMolecules
```

RECRUITMENT AND ACTIVITY OF Plcg2

In addition to Plcg1, we also include Plcg2 in the library because isoform-specific differences between these two proteins have been found in FcεRI signaling. It has been observed that phosphorylation and activation of Plcg2 is less sensitive to PI3K inhibition than Plcg1 (62). Thus, we include a mechanism by which Plcg2 is activated by Syk rather than by Btk. However, we note that other studies have found phosphorylation of Plcg2 to be reduced in the absence of Btk (63), suggesting that Btk may act on Plcg2.

```
(31) Lat(Y136~P)+Plcg2(SH2) <->
     Lat(Y136~P!1).Plcg2(SH2!1) kfLatPlcg,
     krLatPlcg
(32) Syk(tSH2!+,PTK)+Plcg2(SH2!+,Y753~0)
     -> Syk(tSH2!+,PTK!1).Plcg2(SH2!+,
     Y753~0!1) kfSykPlcg
(33) Syk(PTK!1).Plcg2(Y753~0!1) -> Syk(PTK)+
     Plcg2(Y753~0) krSykPlcg
(34) Syk(PTK!1).Plcg2(Y753~0!1) -> Syk(PTK)+
     Plcg2(Y753~P) kpSykPlcg
(35) Plcg2(SH2!+,PLC)+PI45P2(headgroup) ->
     Plcg2(SH2!+,PLC!1).PI45P2(headgroup!1)
     kfPlcgPip2
(36) Plcg2(PLC!1).PI45P2(headgroup!1) ->
     Plcg2(PLC)+PI45P2(headgroup) krPlcgPip2
(37) Plcg2(PLC!1,Y753~P).PI45P2(headgroup!1)
     -> Plcg2(PLC,Y753~P)+IP3()+DAG()
     kcPlcg DeleteMolecules
```

Rule 31 represents binding to Lat. Rules 32–34 represent phosphorylation of Plcg2 through a Michaelis–Menten mechanism. Rule 35–37 represent catalyzed hydrolysis of PIP_2.

ACTIVATION OF Inpp5d

The final regulator of lipid signaling explicitly considered in our model is Inpp5d, also known as Ship1, a phosphatase that can be recruited to FcεRI by binding a non-canonical ITAM tyrosine in the β subunit of the receptor (64, 65). Although Inpp5d and Lyn both bind the β subunit, they have preferences for different phosphotyrosines and thus we treat these interactions as non-competitive. Inpp5d dephosphorylates the 5th position of the inositol ring of PtdIns(3,4,5)P_3 to form PtdIns(3,4)P_2. This product of Inpp5d activity can in turn bind the Inpp5d C2 domain (66), forming a positive feedback loop that has an overall negative impact on FcεRI-mediated degranulation. The following rules are used to model binding of Inpp5d to the receptor:

```
(38) Inpp5d(SH2,C2)+Rec(b_Y224~P) ->
     Inpp5d(SH2!1,C2).Rec(b_Y224~P!1)
     kfShipRec
(39) Inpp5d(IPP,C2!+)+Rec(b_Y224~P) ->
     Inpp5d(IPP!1,C2!+).Rec(b_Y224~P!1)
     100*kfShipRec
(40) Inpp5d(SH2!1).Rec(b_Y224~P!1) ->
     Inpp5d(SH2)+Rec(b_Y224~P) krShipRec
```

In the first rule, Inpp5d is cytosolic, because its SH2 and C2 domains are both free and, in the model, these are the only domains that mediate membrane recruitment. In the second rule, Inpp5d is already membrane associated through binding of its C2 domain to PtdIns(3,4)P_2. For this reason, receptor binding occurs more quickly (we assume a 100-fold enhancement). The third rule represents dissociation of Inpp5d from the receptor.

Binding of Inpp5d to PtdIns(3,4)P_2 is modeled similarly, with different rules for membrane-recruited and cytosolic Inpp5d:

```
(41) Inpp5d(SH2,C2)+PI34P2(headgroup) ->
     Inpp5d(SH2,C2!1).PI34P2(headgroup!1)
     kfShipPip2
(42) Inpp5d(SH2!+,C2)+PI34P2(headgroup) ->
     Inpp5d(SH2!+,C2!1).PI34P2(headgroup!1)
     100*kfShipPip2
(43) Inpp5d(C2!1).PI34P2(headgroup!1) ->
     Inpp5d(C2)+PI34P2(headgroup) krShipPip2
```

In the first rule, Inpp5d is cytosolic, whereas in the second rule, it is localized to the membrane through binding of its SH2 domain to the receptor. As above, a 100-fold enhancement is assumed. The third rule represents dissociation.

Inpp5d ACTIVITY

The following rules capture the catalytic activity of Inpp5d:

```
(44) Inpp5d(SH2!+,C2,IPP)+PI345P3
     (headgroup) -> Inpp5d(SH2!+,C2,IPP!1).
     PI345P3(headgroup!1) kfShipPip3
(45) Inpp5d(SH2,C2!+,IPP)+PI345P3
     (headgroup) -> Inpp5d(SH2,C2!+,IPP!1).
     PI345P3(headgroup!1) kfShipPip3
(46) Inpp5d(SH2!+,C2!+,IPP)+PI345P3
     (headgroup) -> Inpp5d(SH2!+,C2!+,
     IPP!1).PI345P3(headgroup!1) kfShipPip3
(47) Inpp5d(IPP!1).PI345P3(headgroup!1) ->
     Inpp5d(IPP)+PI345P3(headgroup)
     krShipPip3
(48) Inpp5d(IPP!1).PI345P3(headgroup!1) ->
     Inpp5d(IPP)+PI34P2(headgroup)
     kdpShipPip3 DeleteMolecules
```

In the rules above, "IPP" represents the catalytic domain of Inpp5d.

ADDITIONAL LIPID REACTIONS

Conversion of PtdIns(3,4,5)P_3 to PtdIns(4,5)P_2 by Pten is considered implicitly as a first-order reaction. Conversions between PtdIns(4,5)P_2 and PtdIns(4)P are modeled similarly.

```
(49) PI345P3(headgroup) -> PI45P2(headgroup)
     kPten DeleteMolecules
(50) PI4P(headgroup) -> PI45P2(headgroup)
     kfP5 DeleteMolecules
(51) PI45P2(headgroup) -> PI4P(headgroup)
     krP5 DeleteMolecules
```

IDENTIFICATION OF NETWORK MOTIFS

It has been hypothesized that relatively simple network motifs with specialized functions play important roles in cellular regulatory systems and that understanding the design principles of these motifs can help us better understand the complex systems in which they are embedded (67, 68). Network motifs, such as feedback loops, have the potential to generate and/or regulate non-linear dynamical behavior (69), which may, for example, enable precise encoding of information about a stimulus (70). We assessed the FcεRI signaling network for the presence of network motifs, and identified motifs from four classes: positive feedback loops, negative feedback loops, incoherent feed-forward loops, and coherent feed-forward loops. Several of the positive and negative feedbacks contribute to regulation of the SFKs Lyn and Fyn, as well as Syk. One positive feedback loop arises because SFKs phosphorylate tyrosine residues in FcεRI, which serve as binding sites that recruit additional Lyn and Fyn molecules. Furthermore, Lyn and Fyn can each *trans* phosphorylate their own activation loop, which enhances catalytic activity. A similar mechanism also activates the kinase Syk. Negative feedback arises because Lyn and Fyn can phosphorylate the adaptor Pag1, which recruits Csk to negatively regulate SFK activity. This set of interactions has been predicted to lead to oscillations in BCR signaling (38).

Other positive feedback loops are involved in regulating lipid metabolism. PI3K generates PIP_3, which recruits Gab2. Gab2 can in turn recruit additional PI3K. An additional positive feedback loop regulates Inpp5d, because it is capable of binding its own product. Inpp5d is also involved in an incoherent feed-forward loop, meaning a process in which two parallel mechanisms have opposite influences on an output. Here, the output is PIP_3. Inpp5d is recruited to FcεRI and dephosphorylates PIP_3. Incoherence arises because FcεRI contributes to activation of PI3K, which generates PIP_3. In this way, opposing influences are exerted on the abundance of PIP_3 upon stimulation of FcεRI signaling. Such circuitry has been hypothesized to be involved in adaptation, the capacity of a system to respond to an input and then reset itself to a pre-stimulated state (71). Thus, PIP_3 level may be raised and then lowered after a period of FcεRI stimulation, with Inpp5d-mediated positive feedback reinforcing negative regulation over time.

Finally, we identified a pair of coherent feed-forward loops stemming from the adaptor Lat. In a coherent feed-forward loop, two processes exert the same influence (either positive or negative) on an output. In each of the feed-forward loops of interest here, both processes in the network motif have a positive influence on Plcg1 activity. In the first feed-forward loop, Lat recruits Plcg1 via one of its phosphotyrosines. Other Lat phosphotyrosines are involved in assembly of a signaling complex that ultimately recruits PI3K. The product of PI3K, PIP_3, binds the kinase Btk, which phosphorylates Plcg1 at an activating site. In the second feed-forward loop, Lat contributes to Plcg1 recruitment through direct binding as well as through recruitment of another adaptor, Lcp2. What function could be achieved by these (overlapping) feed-forward loops? In transcriptional regulatory networks, it has been found that feed-forward loops can act as sign-sensitive delay elements, meaning that they enable rapid responses to changes in an input in one direction, and slow responses to changes in the input in the opposite direction (72, 73). Thus, the feed-forward loops initiated by Lat may influence the timing of Plcg1 activation and deactivation after increases or decreases in, for example, upstream receptor phosphorylation.

It is worth noting that Plcg1 and PI3K act on the same substrate, PIP_2. Thus, although PI3K can positively influence Plcg1, these two enzymes also compete with one another and could together deplete available PIP_2, assuming both access the same lipid pool. In this way, the feed-forward loop may be self-limiting. For example, if Plcg1 causes rapid conversion of PIP_2 to IP_3, less PIP_2 would be available to PI3K and as a result, less PIP_3 would be generated and the impact of the feed-forward loop would be reduced. The strength of the feed-forward loop would also be influenced by the rate of production of PIP_2 by specific lipid kinases and phosphatases. A caveat is that Plcg1 and PI3K may act on spatially distinct lipid pools, which PIP_2 has been found to exist in (74). These factors are not immediately evident from examination of isolated circuitry. This example highlights the importance of considering broader context and physical parameters (e.g., concentrations and binding affinities) in assessment of network motif functionality.

SENSITIVITY OF PHOSPHOLIPID METABOLISM TO PROTEIN TYROSINE KINASE ACTIVATION

We next used our rule library to develop models for investigation of signaling dynamics. We focused on the adaptor protein Lat, which is known for its role as a signaling hub in both T cells and mast cells (44). This role arises in large part from its capacity to recruit multiple adaptors and enzymes that regulate lipid metabolism and production of second messengers. Most of Lat's interactions depend on prior Lat phosphorylation, which is catalyzed primarily by Syk. However, studies of FcεRI signaling induced by structurally defined antigens have revealed that not all "downstream" events are equally dependent on Lat phosphorylation. Specifically, a panel of rigid antigens, composed of haptenated DNA sequences and differing in the distance between DNP hapten groups, was evaluated for the ability to induce phosphorylation of signaling proteins, Ca^{2+} mobilization, and degranulation. It was found that phosphorylation of FcεRI and Lat, as well as store-operated Ca^{2+} entry and degranulation, were strongly dependent on hapten spacing, with the shortest spacing examined associated with the strongest responses. In contrast, it was also found that release of Ca^{2+} from intracellular stores did not show as strong a dependence on the distance between hapten sites (19). Given that Ca^{2+} release is thought to occur as a result of activities of proteins

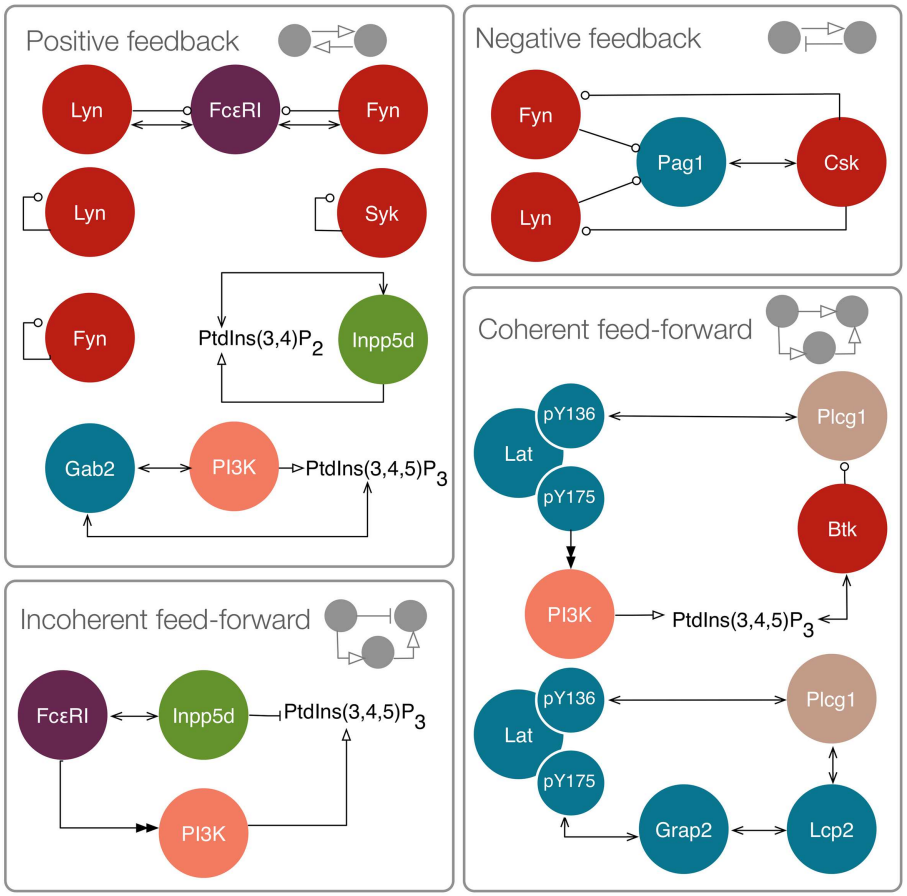


FIGURE 4 | Motifs in the FcεRI signaling network. Positive feedbacks include interactions between FcεRI and Lyn and Fyn, because Lyn and Fyn catalyze phosphorylation of additional binding sites for these kinases. Lyn, Fyn, and Syk are subject to *trans* autophosphorylation at activating sites. Inpp5d binds its own product. Gab2 recruits PI3K, which generates PIP_3, which can recruit additional Gab2. Negative feedback includes inhibition of Lyn and Fyn by Csk. Incoherent feed-forward includes FcεRI stimulation leading to activation of both PI3K and Inpp5d, which exert opposing influences on PIP_3 level. Coherent feed-forwards include recruitment and activation of Plcg1, and recruitment of Plcg1 through two pathways.

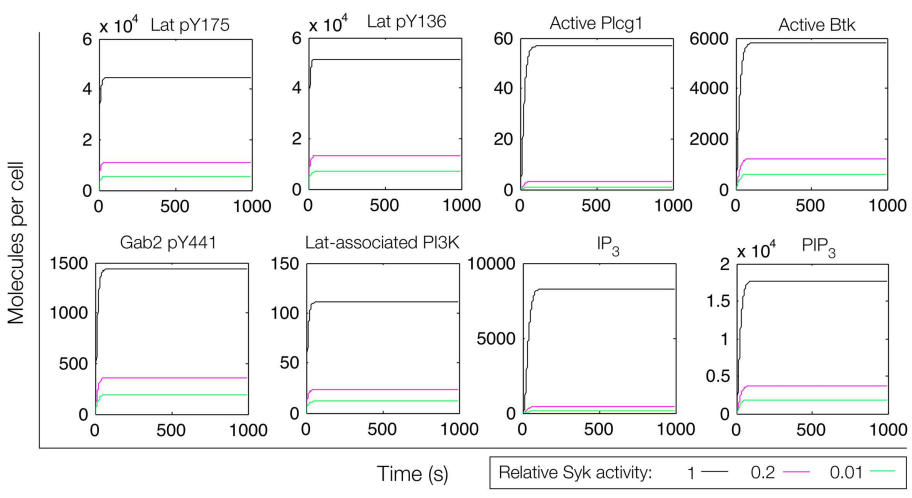


FIGURE 5 | Simulation of a model of the feed-forward loop connecting Lat to IP_3 production. Different color lines indicate different relative levels of Syk activity. In these simulations, Syk activity was set at the indicated level and held constant.

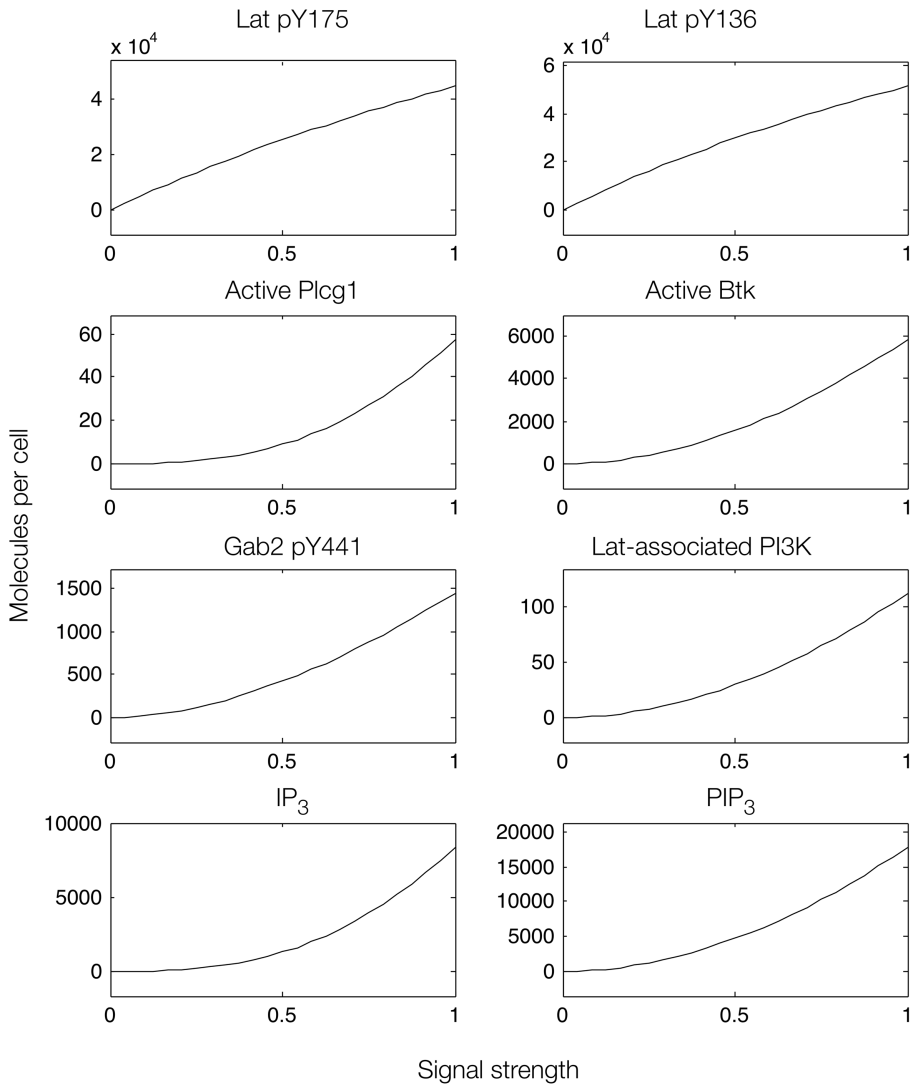


FIGURE 6 | Steady-state dose–response curves, which were found by simulation of the feed-forward loop connecting Lat to IP_3 production. The differences in phosphorylation level between the two phosphorylation sites in Lat (top panels) results from different affinities of the binding partners that interact with each site. Active Plcg1 is taken to be Plcg1 that is both recruited to Lat and phosphorylated, and active Btk is taken to be Btk recruited to PIP_3.

that depend on Lat, how can this apparent uncoupling between Lat phosphorylation and Ca^{2+} mobilization be explained?

We hypothesize that compensatory mechanisms mediated by Fyn and Gab2 (58) are involved in this phenomenon. Gab2 can be phosphorylated by Fyn, and can then recruit PI3K. As discussed above, production of PIP_3 by PI3K contributes to activation of Plcg1. A product of Plcg1 is IP_3, which induces release of Ca^{2+} from intracellular stores. Thus, if Gab2 recruitment and activation is robust to differences in Lat phosphorylation level, then Gab2 may open an avenue by which Ca^{2+} mobilization could escape control of Lat. We used our rule library to build models to determine if Gab2 could potentially enable Ca^{2+} mobilization when Lat phosphorylation is diminished.

We first considered a model in which Syk and Fyn were independent inputs. Our initial model (File S2 in Supplementary Material) essentially consists of the first coherent feed-forward loop shown in **Figure 4**: lat recruits Plcg1, as well as PI3K through Gab2 and Grb2. Btk is recruited to PIP_3 and activates Plcg1 through phosphorylation. Fyn participates by phosphorylating Gab2. To model the differences between antigens observed to induce the most and least Lat phosphorylation, we considered different levels of active Syk consistent with the approximately fourfold difference in Lat phosphorylation observed experimentally (19). The level of active Fyn was kept constant. Simulations of this model revealed that differences in Lat phosphorylation level were maintained or amplified in downstream events. According to the model, a decrease in Lat phosphorylation (arising from lowered Syk activity) causes at least proportionate decreases in the levels of activated Plcg1, activated Btk, Lat-associated PI3K, PIP_3, and IP_3 (**Figure 5**). We also considered a scenario in which activity of Syk and Fyn are both controlled

by the magnitude of an input signal, which may be a more realistic scenario because both kinases are recruited to phosphorylated receptors. We varied the strength of this signal and evaluated the resulting steady-state levels of outputs (**Figure 6**). Consistent with results from the first scenario, decreased signal strength led to decreased Lat phosphorylation, and was accompanied by even steeper decreases in activation of other signaling molecules. Thus, the interactions included in this model are insufficient to explain the experimental observation of Ca^{2+} mobilization in the absence of strong Lat phosphorylation.

In an extension of the initial model (File S3 in Supplementary Material), we incorporated additional interactions from the rule library, those responsible for the positive feedback involving Gab2 interaction with PIP_3 (see **Figure 4**). We reasoned that, with the addition of these interactions, once PIP_3 production is initiated, PIP_3 production may become self-sustaining, because PIP_3 is able to recruit Gab2 to the plasma membrane, which in turn is able to recruit PI3K. Simulated time courses with the same level of active Fyn and different levels of active Syk, as in **Figure 5**, are shown in **Figure 7**. These results indicate that certain signaling readouts downstream of Lat are buffered against reduced Lat phosphorylation. For example, there is less than a fourfold difference in peak IP_3 levels between the conditions of high (black line) and intermediate (magenta line) Lat phosphorylation. In contrast, the model without Gab2-mediated positive feedback predicted a greater than 100-fold difference.

To further investigate the role of positive feedback, we modulated an input signal controlling both Fyn and Syk activity, as in **Figure 6**. Steady-state simulation results from this model are shown in **Figure 8**, which differ from those obtained with the first model. First, the total numbers of signaling molecules in activated forms are greater than for the case without feedback, as long as the signal strength is above a certain level. Second, within certain input ranges, the model shows bistability, i.e., existence of two stable steady states, as indicated by signal strength values that correspond to more than one steady-state output value. Bistability has also been characterized in TCR signaling (75, 76) and BCR signaling (38, 77). Third, we found that certain signaling readouts downstream of Lat are now buffered against reduced Lat phosphorylation (**Figure 7**), decreasing less sharply when signal is reduced. Together, these results suggest that Gab2-mediated positive feedback may enable committed, all-or-none decisions that lead to high levels of IP_3 as long as Lat phosphorylation is above a threshold. When input level falls below this threshold, positive feedback is unable to enhance IP_3 production (**Figure 7**). Thus, some amount of PIP_3 must be generated through Lat-dependent mechanisms before the Fyn/Gab2 pathway can contribute to production of IP_3.

We also considered how positive feedback affects the dynamics of signaling. We calculated the rise time for IP_3 at different input levels as predicted by the models with and without positive feedback. We found that positive feedback caused IP_3 level to reach its steady state more slowly (**Figure 9A**). Rise time for the model with positive feedback peaked in the bistable region, where the system transitions from a low steady state to a higher steady state (**Figure 9B**). The slower rise in IP_3 level qualitatively mimics the experimentally observed dynamics of Ca^{2+} release from stores caused by antigens that induce low levels of Lat phosphorylation (19). These same antigens induce minimal store-operated calcium entry (SOCE) and minimal degranulation, which suggests that SOCE may be sensitive to the kinetics of IP_3 production.

There are several experimental tests that could be pursued to evaluate the role of Gab2-mediated positive feedback. One predicted effect of the feedback loop is bistability of several signaling readouts (**Figure 8**), including PIP_3. Testing for bistability usually benefits from single-cell measurements, because cell-to-cell variability may result in different cells having different bifurcation points. At the single-cell level, PIP_3 production can be monitored using PH domain constructs (78). When the strength of an input signal, such as ligand-induced receptor aggregation, crosses a threshold level, the quantity of PIP_3 is expected to increase dramatically in a switch-like manner. Another characteristic arising from bistability is hysteresis, meaning history dependence. As

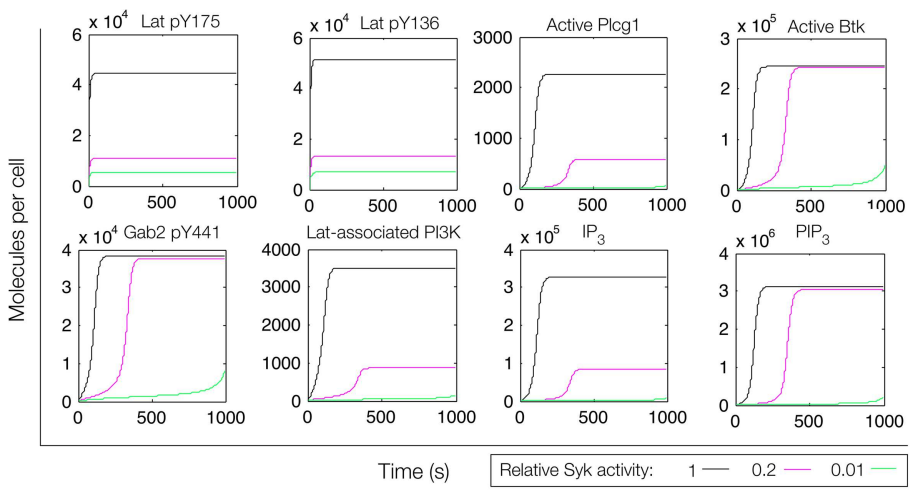


FIGURE 7 | Simulation of a model of the feed-forward loop connecting Lat to IP_3 production with consideration of a Gab2-mediated positive feedback loop. Different color lines indicate different relative levels of Syk activity.

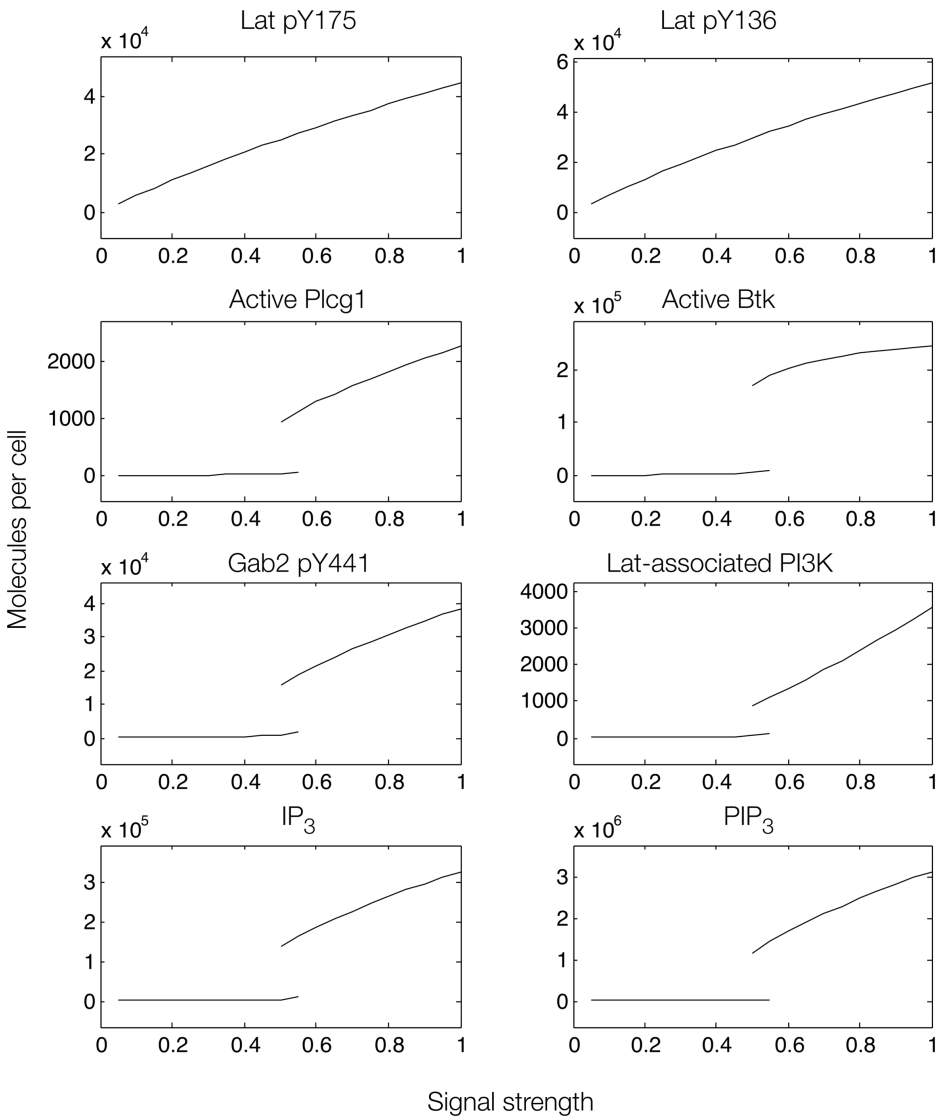


FIGURE 8 | Steady-state dose–response curves, which were found by simulation of the feed-forward loop connecting Lat to IP_3 production when Gab2-mediated positive feedback is considered. Each plot is a bifurcation diagram; the bifurcation parameter is signal strength, which governs the rate of production of active Syk and Fyn. Only stable steady states are shown. As can be seen, the model predicts the possibility of bistability.

signal strength is reduced from a high level [e.g., by breaking up receptor aggregates with a monovalent hapten (34)], PIP_3 level is expected to switch back to a low state. However, this switch is predicted to occur at a lower input level than that required to induce a transition from low signaling to high signaling. Controlling input level would require an understanding of how ligand dose relates to receptor aggregation, which can be obtained with a model for ligand–receptor interactions (79).

A second approach would involve disruption of the Gab2 feedback loop, which would be expected to increase sensitivity to Lat phosphorylation. Mutation of the Gab2 PH domain, which binds PIP_3 and is therefore a key component of the feedback, would be expected to inhibit Ca^{2+} mobilization. However, such manipulation of endogenous Gab2 would be technically challenging, making this strategy potentially difficult to implement. An alternative approach would be to knock down either Gab2 or Fyn, which would be predicted to similarly inhibit Ca^{2+} mobilization.

CONCLUSION

As a step toward systems-level understanding of FcεRI signaling, we have synthesized information about a relatively large number of interactions and proteins into a formalized interaction library. This library consists of executable rules that can be used to extend existing models and to build new models. The rules are annotated with information from the primary literature, thereby facilitating reuse of information. The rules are also visualized to illustrate the scope and detail of the library's contents. Analysis of the library reveals multiple feedback and

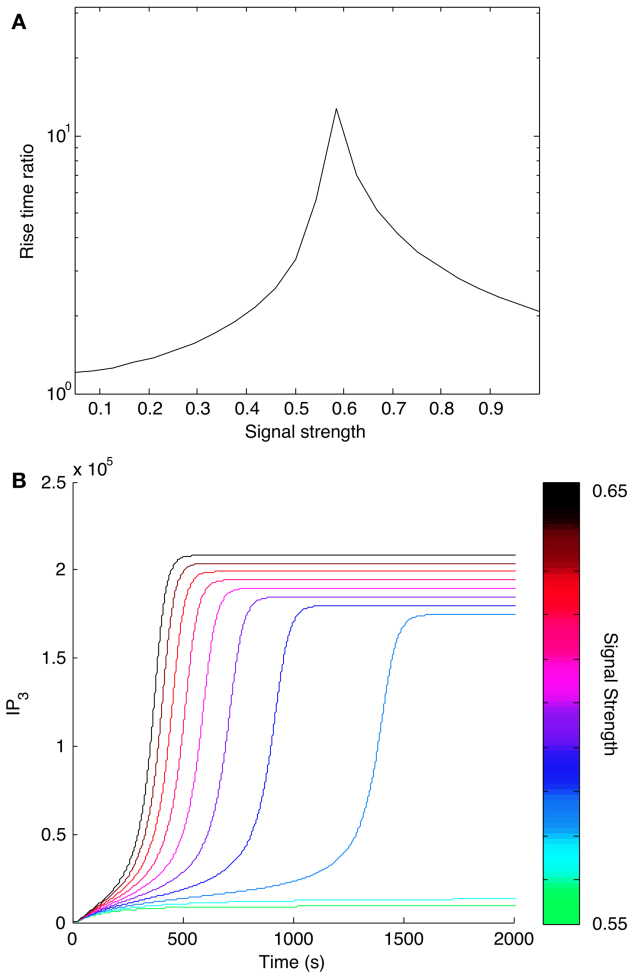


FIGURE 9 | Effect of positive feedback on signaling dynamics. **(A)** Rise time for IP_3 synthesis was calculated as the time needed to reach 95% of the final steady-state level. Rise times were calculated for different levels of input, or signal strength. Rise times for the model with positive feedback were divided by rise times for the model without positive feedback and plotted against corresponding input level. All indicated rise time ratios are greater than one, meaning that the model with positive feedback takes more time to reach its final steady state. **(B)** Time courses for IP_3 production in a narrow range of input levels surrounding the peak shown in **(A)**. The input level corresponding to each curve is indicated with the color bar at the right.

feed-forward loops in FcεRI signaling, the behavior of which can be investigated quantitatively through simulation and complementary quantitative experiments. We used the library to model events involved in phosphoinositide metabolism at different levels of Syk activity, and found that a Gab2-mediated positive feedback can compensate for reduced Lat phosphorylation, which provides a potential explanation for how antigens that induce dramatically different levels of Lat phosphorylation can induce similar Ca^{2+} fluxes (19).

We anticipate that the approach presented here will have several potential applications in linking computational and experimental investigations of cellular information processing. First, a library of rules could be used to build a model of broad scope and site-specific detail for use in analysis of multiplexed, high-resolution data, such as proteomic measurements of site-specific post-translational modifications (80). Currently, such data are often analyzed using clustering, enrichment analysis, and other techniques that reveal trends in dynamics and functions of detected proteins (81), but that do not necessarily provide a concrete picture of the mechanisms at work. Modeling will enable us to better leverage information about mechanisms and physical parameters, complementing current analysis techniques. A combination of modeling and quantitative high-throughput experimentation could, for example, be used to characterize the distribution of signaling complexes that can be nucleated by Lat. Binding partners of Lat have shared binding sites and a range of affinities (49). To understand how binding of these proteins is balanced, it would be necessary to measure binding affinities of SH2 domains to each phosphosite (82) and to quantify protein copy numbers (83). A model could then be used to integrate such data and determine the expected distribution of signaling complexes.

Second, rule libraries could facilitate the extension of models by increments. The benefit of such an approach is that a model of an idealized network motif (71, 84) could be extended piece by piece to form a more complete representation of the motif's context. Studies of such models could reveal how well the predicted behavior of an isolated motif is maintained when additional interactions are considered, and what complicating factors may need to be taken into account in experimental assessments of motif function or in synthetic biology efforts aimed at engineering regulatory systems on the basis of network motif design principles.

Finally, rule libraries may help address problems in knowledge engineering, i.e., the task of gathering, organizing, and interpreting large quantities of information. Rule-based models have already been annotated using interactive wikis (85, 86), which could open the door to community-based model development and curation, making it easier to assemble and assess data for model building. Furthermore, a widely used approach in knowledge engineering is natural language processing (NLP), the automated derivation of information from text. A major bioinformatics goal of NLP is to extract networks and quantitative models from the primary biomedical literature (87). A limiting factor in this task is the availability of "gold standard" networks against which to compare an automatically constructed network or model, which is necessary to assess the performance of network/model construction algorithms. For many biological systems, reliable network representations and models are non-existent. Furthermore, even when a reliable network is available, it may be in a format (e.g., ordinary differential equations) that does not map to underlying interactions in a clear manner. This problem is addressed by a rule library, because rules not only serve as the basis for simulations but also provide precise, human- and machine-readable representations of biomolecular interactions. NLP could aid in library construction through automatic extraction of rules from the literature. As information about cell signaling systems continues to expand, we anticipate that formalization and synthesis of knowledge will

become increasingly important for informing hypotheses, making quantitative predictions, and elucidating systems-level properties of cellular regulatory systems.

ACKNOWLEDGMENTS

This work was supported by NIH grants R01 AI018306 and P50GM085273 and by US Department of Energy Contract DE-AC52-06NA25396 through the Los Alamos Center for Non-linear Studies (CNLS) and the Laboratory-Directed Research and Development (LDRD) Program. We thank Leonard A. Harris for technical advice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00172/abstract

REFERENCES

1. Smith-Garvin JE, Koretzky GA, Jordan MS. T cell activation. *Annu Rev Immunol* (2009) **27**:591–619. doi:10.1146/annurev.immunol.021908.132706
2. Packard TM, Cambier JC. B lymphocyte antigen receptor signaling: initiation, amplification, and regulation. *F1000Prime Rep* (2013) **5**:40. doi:10.12703/P5-40
3. Kraft S, Kinet JP. New developments in FcεRI regulation, function and inhibition. *Nat Rev Immunol* (2007) **7**:365–78. doi:10.1038/nri2072
4. Wollman R, Meyer T. Coordinated oscillations in cortical actin and Ca^{2+} correlate with cycles of vesicle secretion. *Nat Cell Biol* (2012) **14**:1261–9. doi:10.1038/ncb2614
5. Wu M, Wu X, De Camilli P. Calcium oscillations-coupled conversion of actin travelling waves to standing oscillations. *Proc Natl Acad Sci U S A* (2013) **110**:1339–44. doi:10.1073/pnas.1221538110
6. Fraser I, Germain R. Navigating the network: signaling cross-talk in hematopoietic cells. *Nat Immunol* (2009) **10**:327–31. doi:10.1038/ni.1711
7. Kohn K, Aladjem M, Weinstein J, Pommier Y. Network architecture of signaling from uncoupled helicase-polymerase to cell cycle checkpoints and trans-lesion DNA synthesis. *Cell Cycle* (2009) **8**:2281–99. doi:10.4161/cc.8.14.9102
8. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The systems biology graphical notation. *Nat Biotechnol* (2009) **27**:735–41. doi:10.1038/nbt.1558
9. Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, et al. A comprehensive map of the mTOR signaling network. *Mol Syst Biol* (2010) **6**:453. doi:10.1038/msb.2010.108
10. Chylek LA, Hu B, Blinov ML, Emonet T, Faeder JR, Goldstein B, et al. Guidelines for visualizing and annotating rule-based models. *Mol Biosyst* (2011) **7**:2779–95. doi:10.1039/c1mb05077j
11. Tiger C, Krause F, Cedersund G, Palmér R, Klipp E, Hohmann S, et al. A framework for mapping, visualisation and automatic model creation of signal-transduction networks. *Mol Syst Biol* (2012) **8**:578. doi:10.1038/msb.2012.12
12. Kirouac D, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger D, Sorger P. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst Biol* (2012) **6**:29. doi:10.1186/1752-0509-6-29
13. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser ID. Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol* (2011) **29**:527–85. doi:10.1146/annurev-immunol-030409-101317
14. Chylek LA, Harris LA, Tung C-S, Faeder JR, Lopez CF, Hlavacek WS. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley Interdiscip Rev Syst Biol Med* (2014) **6**:13–36. doi:10.1002/wsbm.1245
15. Chylek LA. Decoding the language of phosphorylation site dynamics. *Sci Signal* (2013) **6**:jc2. doi:10.1126/scisignal.2004061
16. Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W. Rules for modeling signal-transduction systems. *Sci STKE* (2006) **2006**:re6. doi:10.1126/stke.3442006re6
17. Chylek LA, Stites EC, Posner RG, Hlavacek WS. Innovations of the rule-based modeling approach. In: Prokop A, Csukás B, editors. *Systems Biology: Integrative Biology and Simulation Tools* (Vol. 1). Dordrecht: Springer (2013). p. 273–300.
18. Paar J, Harris N, Holowka D, Baird B. Bivalent ligands with rigid double-stranded DNA spacers reveal structural constraints on signaling by FcεRI. *J Immunol* (2002) **169**:856–64.
19. Sil D, Lee J, Luo D, Holowka D, Baird B. Trivalent ligands with rigid DNA spacers reveal structural requirements for IgE receptor signaling in RBL mast cells. *ACS Chem Biol* (2007) **2**:674–84. doi:10.1021/cb7001472
20. Posner R, Geng D, Haymore S, Bogert J, Pecht I, Licht A, et al. Trivalent antigens for degranulation of mast cells. *Org Lett* (2007) **9**:3551–4. doi:10.1021/ol071175h
21. Goldstein B, Faeder JR, Hlavacek WS, Blinov ML, Redondo A, Wofsy C. Modeling the early signaling events mediated by FcεRI. *Mol Immunol* (2002) **38**:1213–9. doi:10.1016/S0161-5890(02)00066-4
22. Faeder JR, Hlavacek WS, Reischl I, Blinov ML, Metzger H, Redondo A, et al. Investigation of early events in FcεRI-mediated signaling using a detailed mathematical model. *J Immunol* (2003) **170**:3769–81.
23. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS. Combinatorial complexity and dynamical restriction of network flows in signal transduction. *Syst Biol (Stevenage)* (2005) **2**:5–15. doi:10.1049/sb:20045031
24. Nag A, Monine MI, Faeder JR, Goldstein B. Aggregation of membrane proteins by cytosolic cross-linkers: theory and simulation of the LAT-Grb2-SOS1 system. *Biophys J* (2009) **96**:2604–23. doi:10.1016/j.bpj.2009.01.019
25. Nag A, Monine M, Perelson A, Goldstein B. Modeling and simulation of aggregation of membrane protein LAT with molecular variability in the number of binding sites for cytosolic Grb2-SOS1-Grb2. *PLoS One* (2012) **7**:e28758. doi:10.1371/journal.pone.0028758
26. Nag A, Faeder JR, Goldstein B. Shaping the response: the role of FcεRI and Syk expression levels in mast cell signaling. *IET Syst Biol* (2010) **33**(4–47):334–47. doi:10.1049/iet-syb.2010.0006
27. Nag A, Monine MI, Blinov ML, Goldstein B. A detailed mathematical model predicts that serial engagement of IgE-Fc epsilon RI complexes can enhance Syk activation in mast cells. *J Immunol* (2010) **185**:3268–76. doi:10.4049/jimmunol.1000326
28. Cao L, Yu K, Banh C, Nguyen V, Ritz A, Raphael B, et al. Quantitative time-resolved phosphoproteomic analysis of mast cell signaling. *J Immunol* (2007) **179**:5864–76.
29. Gilfillan A, Tkaczyk C. Integrated signalling pathways for mast-cell activation. *Nat Rev Immunol* (2006) **16**:218–30. doi:10.1038/nri1782
30. Faeder JR, Blinov ML, Hlavacek WS. Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol Biol* (2009) **500**:113–67. doi:10.1007/978-1-59745-525-1_5
31. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, et al. Sundials: suite of nonlinear and differential/algebraic equation solvers. *ACM Trans Math Softw* (2005) **31**:363–96. doi:10.1145/1089014.1089020
32. Holowka D, Sil D, Torigoe C, Baird B. Insights into immunoglobulin E receptor signaling from structurally defined ligands. *Immunol Rev* (2007) **217**:269–79. doi:10.1111/j.1600-065X.2007.00517.x
33. Xu K, Goldstein B, Holowka D, Baird B. Kinetics of multivalent antigen DNP-BSA binding to IgE-Fc epsilon RI in relationship to the stimulated tyrosine phosphorylation of Fc epsilon RI. *J Immunol* (1998) **160**:3225–35.
34. Shelby SA, Holowka D, Baird B, Veatch SL. Distinct stages of stimulated FcεRI receptor clustering and immobilization are identified through superresolution imaging. *Biophys J* (2013) **105**:2343–54. doi:10.1016/j.bpj.2013.09.049
35. Liu Y, Barua D, Liu P, Wilson BS, Oliver JM, Hlavacek WS, et al. Single-cell measurements of IgE-mediated FcεRI signaling using an integrated microfluidic platform. *PLoS One* (2013) **8**:e60159. doi:10.1371/journal.pone.0060159
36. Young RM, Holowka D, Baird B. A lipid raft environment enhances Lyn kinase activity by protecting the active site tyrosine from dephosphorylation. *J Biol Chem* (2003) **278**:20746–52. doi:10.1074/jbc.M211402200
37. Jouvin MH, Adamczewski M, Numerof R, Letourneur O, Vallé A, Kinet JP. Differential control of the tyrosine kinases Lyn and Syk by the two signaling chains of the high affinity immunoglobulin E receptor. *J Biol Chem* (1994) **269**:5918–25.
38. Barua D, Hlavacek WS, Lipniacki T. A computational model for early events in B cell antigen receptor signaling: analysis of the roles of Lyn and Fyn. *J Immunol* (2012) **189**:646–58. doi:10.4049/jimmunol.1102003

39. Consortium TU. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* (2014) **42**:D191–8. doi:10.1093/nar/gkt1140
40. Holowka D, Calloway N, Cohen R, Gadi D, Lee J, Smith NL, et al. Roles for Ca^{2+} mobilization and its regulation in mast cell functions. *Front Immunol* (2012) **3**:104. doi:10.3389/fimmu.2012.00104
41. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* (2010) **11**:R3. doi:10.1186/gb-2010-11-1-r3
42. Zhang W, Trible RP, Samelson LE. LAT palmitoylation: its essential role in membrane microdomain targeting and tyrosine phosphorylation during T cell activation. *Immunity* (1998) **9**:239–46. doi:10.1016/S1074-7613(00)80606-8
43. Zhang W, Sloan-Lancaster J, Kitchen J, Trible RP, Samelson LE. LAT: the ZAP-70 tyrosine kinase substrate that links T cell receptor to cellular activation. *Cell* (1998) **92**:83–92. doi:10.1016/S0092-8674(00)80901-0
44. Saitoh S, Arudchandran R, Manetz T, Zhang W, Sommers C, Love P, et al. Lat is essential for FcεRI-mediated mast cell activation. *Immunity* (2000) **12**:525–35. doi:10.1016/S1074-7613(00)80204-6
45. Das R, Hammond S, Holowka D, Baird B. Real-time cross-correlation image analysis of early events in IgE receptor signaling. *Biophys J* (2008) **94**:4996–5008. doi:10.1529/biophysj.107.105502
46. Veatch SL, Chiang EN, Sengupta P, Holowka DA, Baird BA. Quantitative nanoscale analysis of IgE-FcεRI clustering and coupling to early signaling proteins. *J Phys Chem B* (2012) **116**:6923–35. doi:10.1021/jp300197p
47. Barua D, Goldstein B. A mechanistic model of early FcεRI signaling: lipid rafts and the question of protection from dephosphorylation. *PLoS One* (2012) **7**:e51669. doi:10.1371/journal.pone.0051669
48. Zhang J, Billingsley ML, Kincaid RL, Siraganian RP. Phosphorylation of Syk activation loop tyrosines is essential for Syk function. An in vivo study using a specific anti-Syk activation loop phosphotyrosine antibody. *J Biol Chem* (2000) **275**:35442–7. doi:10.1074/jbc.M004549200
49. Houtman JC, Higashimoto Y, Dimasi N, Cho S, Yamaguchi H, Bowden B, et al. Binding specificity of multiprotein signaling complexes is determined by both cooperative interactions and affinity preferences. *Biochemistry* (2004) **43**:4170–8. doi:10.1021/bi0357311
50. Jang IK, Zhang J, Gu H. Grb2, a simple adapter with complex roles in lymphocyte development, function, and signaling. *Immunol Rev* (2009) **232**:150–9. doi:10.1111/j.1600-065X.2009.00842.x
51. Cho S, Velikovsky C, Swaminathan CP, Houtman JC, Samelson LE, Mariuzza RA. Structural basis for differential recognition of tyrosine-phosphorylated sites in the linker for activation of T cells (LAT) by the adaptor gads. *EMBO J* (2004) **23**:1441–51. doi:10.1038/sj.emboj.7600168
52. Seet BT, Berry DM, Maltzman JS, Shabason J, Raina M, Koretzky GA, et al. Efficient T-cell receptor signaling requires a high-affinity interaction between the Gads C-SH3 domain and the SLP-76 RxxK motif. *EMBO J* (2007) **26**:678–89. doi:10.1038/sj.emboj.7601535
53. Stauffer TP, Meyer T. Compartmentalized IgE receptor-mediated signal transduction in living cells. *J Cell Biol* (1997) **139**:1447–54. doi:10.1083/jcb.139.6.1447
54. Samelson LE. Signal transduction mediated by the T cell antigen receptor: the role of adapter proteins. *Annu Rev Immunol* (2002) **20**:371–94. doi:10.1146/annurev.immunol.20.092601.111357
55. Yablonski D, Kadlecek T, Weiss A. Identification of a phospholipase C-gamma1 (PLC-gamma1) SH3 domain-binding site in SLP-76 required for T-cell receptor-mediated activation of PLC-gamma1 and NFAT. *Mol Cell Biol* (2001) **21**:4208–18. doi:10.1128/MCB.21.13.4208-4218.2001
56. Lewitzky M, Kardinal C, Gehring NH, Schmidt EK, Konkol B, Eulitz M, et al. The C-terminal SH3 domain of the adapter protein Grb2 binds with high affinity to sequences in Gab1 and SLP-76 which lack the SH3-typical P-x-x-P core motif. *Oncogene* (2001) **20**:1052–62. doi:10.1038/sj.onc.1204202
57. Edmead CE, Fox BC, Stace C, Ktistakis N, Welham MJ. The pleckstrin homology domain of Gab-2 is required for optimal interleukin-3 signalsome-mediated responses. *Cell Signal* (2006) **18**:1147–55. doi:10.1016/j.cellsig.2005.09.002
58. Parravicini V, Gadina M, Kovarova M, Odom S, Gonzalez-Espinosa C, Furumoto Y, et al. Fyn kinase initiates complementary signals required for IgE-dependent mast cell degranulation. *Nat Immunol* (2002) **3**:741–8. doi:10.1038/ni817
59. Gu H, Maeda H, Moon JJ, Lord JD, Yoakim M, Nelson BH, et al. New role for Shc in activation of the phosphatidylinositol 3-kinase/Akt pathway. *Mol Cell Biol* (2000) **20**:7109–20. doi:10.1128/MCB.20.19.7109-7120.2000
60. Qiu Y, Kung HJ. Signaling network of the Btk family kinases. *Oncogene* (2000) **19**:5651–61. doi:10.1038/sj.onc.1203958
61. Oh-hora M, Rao A. Calcium signaling in lymphocytes. *Curr Opin Immunol* (2008) **20**:250–8. doi:10.1016/j.coi.2008.04.004
62. Barker S, Caldwell K, Pfeiffer J, Wilson B. Wortmannin-sensitive phosphorylation, translocation, and activation of PLCγ1, but not PLCγ2, in antigen-stimulated RBL-2H3 mast cells. *Mol Biol Cell* (1998) **9**:483–96. doi:10.1091/mbc.9.2.483
63. Kawakami Y, Kitaura J, Satterthwaite AB, Kato RM, Asai K, Hartman SE, et al. Redundant and opposing functions of two tyrosine kinases, Btk and Lyn, in mast cell activation. *J Immunol* (2000) **165**:1210–9.
64. Kimura T, Sakamoto H, Appella E, Siraganian RP. The negative signaling molecule SH2 domain-containing inositol-polyphosphate 5-phosphatase (SHIP) binds to the tyrosine-phosphorylated beta subunit of the high affinity IgE receptor. *J Biol Chem* (1997) **272**:13991–6. doi:10.1074/jbc.272.21.13991
65. On M, Billingsley JM, Jouvin MH, Kinet JP. Molecular dissection of the FcRβ signaling amplifier. *J Biol Chem* (2004) **279**:45782–90. doi:10.1074/jbc.M404890200
66. Condé C, Gloire G, Piette J. Enzymatic and non-enzymatic activities of SHIP-1 in signal transduction and cancer. *Biochem Pharmacol* (2011) **82**:1320–34. doi:10.1016/j.bcp.2011.05.031
67. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science* (2002) **298**:842–7. doi:10.1126/science.298.5594.824
68. Lim WA, Lee C, Tang C. Design principles of regulatory networks: searching for the molecular algorithms of the cell. *Mol Cell* (2013) **49**:202–12. doi:10.1016/j.molcel.2012.12.020
69. Tyson J, Novak B. Functional motifs in biochemical reaction networks. *Annu Rev Phys Chem* (2010) **61**:219–40. doi:10.1146/annurev.physchem.012809.103457
70. Dolmetsch R, Xu K, Lewis R. Calcium oscillations increase the efficiency and specificity of gene expression. *Nature* (1998) **392**:933–6. doi:10.1038/31960
71. Ma W, Trusina A, El-Samad H, Lim W, Tang C. Defining network topologies that can achieve biochemical adaptation. *Cell* (2009) **138**:760–73. doi:10.1016/j.cell.2009.06.013
72. Mangan S, Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* (2003) **100**:11980–5. doi:10.1073/pnas.2133841100
73. Mangan S, Zaslaver A, Alon U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* (2003) **334**:197–204. doi:10.1016/j.jmb.2003.09.049
74. Calloway N, Owens T, Corwith K, Rodgers W, Holowka D, Baird B. Stimulated association of STIM1 and Orai1 is regulated by the balance of $Ptdins(4,5)P_2$ between distinct membrane pools. *J Cell Sci* (2011) **124**:2602–10. doi:10.1242/jcs.084178
75. Lipniacki T, Hat B, Faeder JR, Hlavacek WS. Stochastic effects and bistability in T cell receptor signaling. *J Theor Biol* (2008) **254**:110–22. doi:10.1016/j.jtbi.2008.05.001
76. Das J, Ho M, Zikherman J, Govern C, Yang M, Weiss A, et al. Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell* (2009) **136**:337–51. doi:10.1016/j.cell.2008.11.051
77. Mukherjee S, Zhu J, Zikherman J, Parameswaran R, Kadlecek TA, Wang Q, et al. Monovalent and multivalent ligation of the B cell receptor exhibit differential dependence upon Syk and Src family kinases. *Sci Signal* (2013) **6**:ra1. doi:10.1126/scisignal.2003220
78. Halet G. Imaging phosphoinositide dynamics using GFP-tagged protein domains. *Biol Cell* (2005) **97**:501–18. doi:10.1042/BC20040080
79. Monine MI, Posner RG, Savage PB, Faeder JR, Hlavacek WS. Modeling multivalent ligand-receptor interactions with steric constraints on configurations of cell-surface receptor aggregates. *Biophys J* (2010) **98**:48–56. doi:10.1016/j.bpj.2009.09.043
80. Rigbolt K, Blagoev B. Quantitative phosphoproteomics to characterize signaling networks. *Semin Cell Dev Biol* (2012) **23**:863–71. doi:10.1016/j.semcdb.2012.05.006
81. Kumar C, Mann M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett* (2009) **583**:1703–12. doi:10.1016/j.febslet.2009.03.035
82. Hause RJJ, Leung KK, Barkinge JL, Ciaccio M, Chuu C, Jones RB. Comprehensive binary interaction mapping of SH2 domains via fluorescence polarization

reveals novel functional diversification of ErbB receptors. *PLoS One* (2012) **7**:e44471. doi:10.1371/journal.pone.0044471
83. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* (2011) **7**:578. doi:10.1038/msb.2011.81
84. Rowland M, Fontana W, Deeds E. Crosstalk and competition in signaling networks. *Biophys J* (2012) **3**:2389–98. doi:10.1016/j.bpj.2012.10.006
85. Thomson TM, Benjamin KR, Bush A, Love T, Pincus D, Resnekov O, et al. Scaffold number in yeast signaling system sets tradeoff between system output and dynamic range. *Proc Natl Acad Sci U S A* (2011) **108**:20265–70. doi:10.1073/pnas.1004042108
86. Barua D, Hlavacek WS. Modeling the effect of APC truncation on destruction complex function in colorectal cancer cells. *PLoS Comput Biol* (2013) **9**:e1003217. doi:10.1371/journal.pcbi.1003217
87. Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges. *Brief Bioinform* (2013). doi:10.1093/bib/bbt006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 December 2013; accepted: 31 March 2014; published online: 15 April 2014. Citation: Chylek LA, Holowka DA, Baird BA and Hlavacek WS (2014) An interaction library for the FcεRI signaling network. Front. Immunol. ***5****:172. doi: 10.3389/fimmu.2014.00172*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2014 Chylek, Holowka, Baird and Hlavacek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
IMMUNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 December 2013
doi: 10.3389/fimmu.2013.00426

Asymmetry in erythroid-myeloid differentiation switch and the role of timing in a binary cell-fate decision

Afnan Alagha*[1] *and Alexey Zaikin*[2]

[1] *Nonlinear Analysis and Applied Mathematics Research Group (NAAM), Department of Mathematics, King Abdulaziz University, Jeddah, Saudi Arabia*
[2] *Department of Mathematics and Institute for Women's Health, University College London, London, UK*

Edited by:
Carmen Molina-Paris, University of Leeds, UK

Reviewed by:
Koji Yasutomo, University of Tokushima, Japan
Maria L. Toribio, Spanish Research Council (CSIC), Spain

***Correspondence:**
Alexey Zaikin, Department of Mathematics and Institute for Women's Health, University College London, Gower Street, London WC1E 6BT, UK
e-mail: alexey.zaikin@ucl.ac.uk

GATA1-PU.1 genetic switch is a paradigmatic genetic switch that governs the differentiation of progenitor cells into two different fates, erythroid and myeloid fates. In terms of dynamical model representation of these fates or lineages corresponds to stable attractor and choosing between the attractors. Small asymmetries and stochasticity intrinsically present in all genetic switches lead to the effect of delayed bifurcation which will change the differentiation result according to the timing of the process and affect the proportion of erythroid versus myeloid cells. We consider the differentiation bifurcation scenario in which there is a symmetry-breaking in the bifurcation diagrams as a result of asymmetry in external signaling. We show that the decision between two alternative cell fates in this structurally symmetric decision circuit can be biased depending on the speed at which the system is forced to go through the decision point. The parameter sweeping speed can also reduce the effect of asymmetry and produce symmetric choice between attractors, or convert the favorable attractor. This conversion may have important contributions to the immune system when the bias is in favor of the attractor which gives rise to non-immune cells.

Keywords: GATA1-PU.1 switch, differentiation, immune cells, pluripotent cells

1. INTRODUCTION

The importance of studying the immune system has attracted mathematicians and biologists to discover more of its features in recent years. One of the mechanisms is to study the genetic networks that control the lineage commitment of hematopoietic stem cells, which produce the full range of blood cells, including the immune cells (1). Many mathematical models have been used to study the differentiation of progenitor cell into erythroid and myeloid lineages based on the expression of lineage-specific transcription factors GATA1 and PU.1, respectively (2, 3). An important question arises in these models about the causes of bifurcation and symmetry-breaking and whether they occur in response to intrinsic cues or extrinsic signals. In fact, the integration of both intrinsic and extrinsic factors has received an extensive attention to elucidate the roles of external signals in cell-fate decision processes, and most importantly its relationship to the production of immune cells (3–6). Another important and interesting factor that can affect the decision of the cell is the speed of external signals or the speed of crossing the critical region (7–9). Remarkably, varying control parameter with time has been studied in many other systems. Ashwin et al. (10) have investigated how the rate of change of a parameter (or input) imposes significant changes in the climate system. It is found that rapid change may force the system to move away from a branch of attractors. This dependence on the rate was referred to as R-tipping. Another more recent study (11) has discovered how the stress response in bacteria is determined by the rate of environmental change. An increase in environmental stress leads to a single uniform pulse of alternative sigma factor σ^B activation, a general stress response pathway, with amplitude depending on the rate at which the stress increased. It is found that faster stresses lead to larger and sharper activation of σ^B, reflecting the fact that the activation process is rate-dependent. A question naturally arises how rate dependent signaling will affect the immune cell-fate selection via a differentiation of progenitor cells. We have studied these phenomena in the most paradigmatic switch responsible for the differentiation of immune cells, the GATA1-PU.1 switch. Moreover, we have considered how the shape of external signals may have an impact in decision-making process. The paper is structured as follows, we review the model of Huang et al. (2) and investigate, in addition to the symmetric scenario, the asymmetric scenario in two ways: (i) under the effect of asymmetric change of parameters; and (ii) under the effect of external signals, using two kinds of signals (see Materials and Methods). Furthermore, we will test the effect of parameter sweeping speed on the distribution of trajectories in the attractors of the dynamical system.

2. MATERIALS AND METHODS

2.1. THE GATA1-PU.1 GENE REGULATORY CIRCUIT

The model of the genetic switch responsible for differentiation contains mutual inhibition and is shown in (**Figure 1A**). The regulatory dynamics can be described by the following form (2):

$$\frac{dX_1}{dt} = \frac{a_1 X_1^n}{r_{a_1}^n + X_1^n} + \frac{b_1 r_{b_1}^n}{r_{b_1}^n + X_2^n} - k_1 X_1 + \sigma_{X_1}\xi_{X_1} \quad (1)$$

$$\frac{dX_2}{dt} = \frac{a_2 X_2^n}{r_{a_2}^n + X_2^n} + \frac{b_2 r_{b_2}^n}{r_{b_2}^n + X_1^n} - k_2 X_2 + \sigma_{X_2}\xi_{X_2} \quad (2)$$

where X_1 and X_2 are the concentrations of two transcription factors GATA1 and PU.1, respectively. These equations model the dynamics of self-activation and cross-inhibition with Hill functions (12). The parameters a_1, a_2 represent self-activation rates, the parameters b_1, b_2 are basal expression rates, k_1, k_2 are deactivation rates, the parameters r's are thresholds at which the inflection point in the Hill function occurs, and n is the Hill coefficient. The first terms of equations (1) and (2) give the contribution from self-activation, while the second terms measure the effect of cross-inhibition on basal activation rates, and the third terms the degradation. To take account of intrinsic gene expression stochasticity, we consider the differential equations (1) and (2) in the Langevin form by adding multiplicative noise terms (the last ones) where ξ_{X_1} and ξ_{X_2} stand for a Gaussian noise and $\sigma_{X_{1,2}}$ depend on $X_{1,2}$ as suggested in Ref. (13). These noise terms model the contribution of intrinsic noise which is unavoidable in biological systems. External cell signaling can be included in the model as follows

$$\frac{dX_1}{dt} = \frac{a_1 S_1 X_1^n}{r_{a_1}^n + X_1^n} + \frac{b_1 r_{b_1}^n}{r_{b_1}^n + X_2^n} - k_1 X_1 \tag{3}$$

$$\frac{dX_2}{dt} = \frac{a_2 S_2 X_2^n}{r_{a_2}^n + X_2^n} + \frac{b_2 r_{b_2}^n}{r_{b_2}^n + X_1^n} - k_2 X_2 \tag{4}$$

where S_1 and S_2 represent external signals to the genetic switch. Here, we are interested in two generic forms of signals:

- *Linear signals*: In this form (7) the external signals may have different rising times but they are equal in the steady state at $S_{max} = 10$ (see **Figure 2A**). For the sake of simplicity we assume that S_1 reaches to the steady state faster than S_2, and thus the rising time T_1 of S_1 is smaller than the rising time T_2 of S_2. They both increase linearly with time according to

$$S_1(t) = \begin{cases} \frac{S_{max}}{T_1} t & if \quad t \le T_1 \\ S_{max} & if \quad t > T_1 \end{cases} \tag{5}$$

$$S_2(t) = \begin{cases} \frac{S_{max}}{T_2} t & if \quad t \le T_2 \\ S_{max} & if \quad t > T_2 \end{cases} \tag{6}$$

The difference between S_1 and S_2 and the maximal difference A (**Figure 2B**) are defined as follows

$$\Delta S(t) = S_1(t) - S_2(t),\ A = max(\Delta S(t)) = S_{max}\left(1 - \frac{T_1}{T_2}\right) \tag{7}$$

- *Adaptation form of signals*: As suggested in Ref. (14) to achieve biochemical adaptation the signals have transient growth stage where they reach to their maxima, and decay stage where they decay and saturate to their steady states (see **Figure 3**). As for the first form, S_1 has a rising time, θ_1, smaller than S_2, θ_2, and the

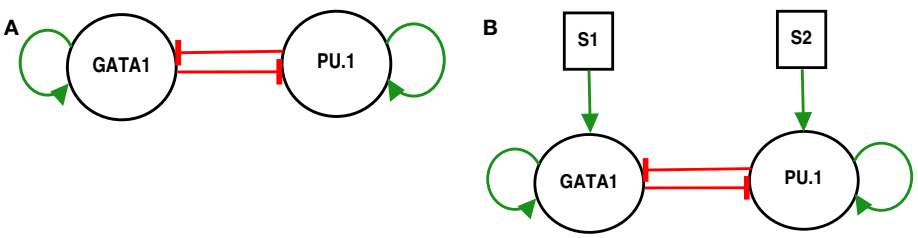


FIGURE 1 | GATA1-PU.1 genetic switch with and without external signals. **(A)** The isolated switch consists of two transcription factors GATA1 and PU.1 that activate themselves while inhibit each other's expression. **(B)** The exposure of the same switch in **(A)** to two external signals S_1 and S_2.

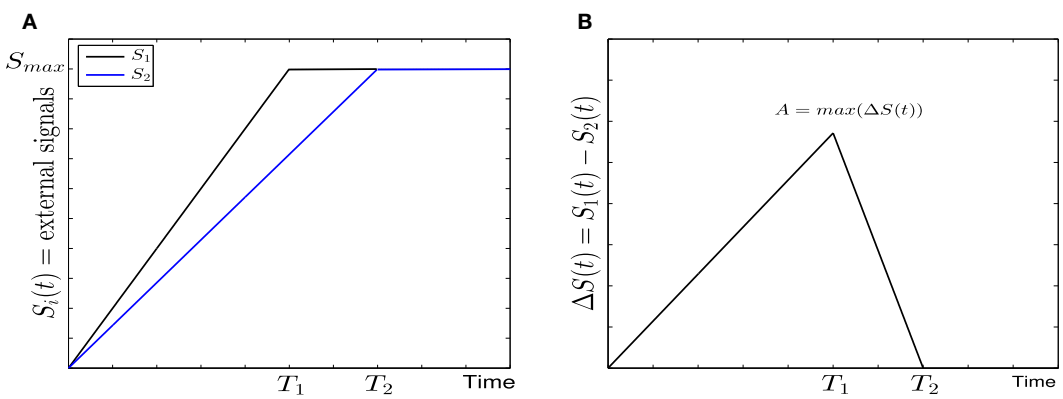


FIGURE 2 | Linear form of external signals in GATA1-PU.1 genetic switch. **(A)** Two external signals S_1 and S_2 with different rising times but equal steady states at $S_{max} = 10$. **(B)** The difference between the external signals with maximal asymmetry at A.

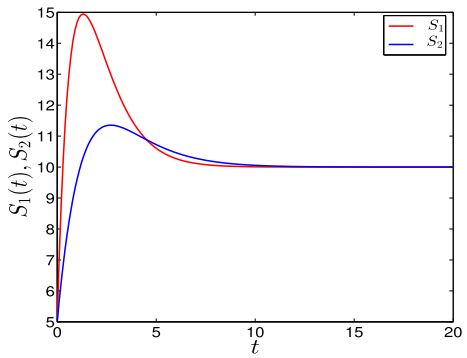


FIGURE 3 | Adaptation form of external signals in GATA1-PU.1 genetic switch. The external signals S_1 and S_2 have different rising times but equal steady states at $v = 10$. Note that at the end of the signaling the system is structurally symmetric.

value of saturation is 10. They have the following form

$$S_1(t) = \frac{h_1}{\theta_1^2} t e^{-\frac{t}{\theta_1}} + \frac{v}{1+e^{-t}} \tag{8}$$

$$S_2(t) = \frac{h_2}{\theta_2^2} t e^{-\frac{t}{\theta_2}} + \frac{v}{1+e^{-t}} \tag{9}$$

where h_1, h_2 control the amplitude of signals, and θ_1, θ_2 are scale parameters. The second terms control the saturation of the signals to the value $v = 10$ (selected value). The maxima occur at $t_{max} = \theta_{1,2}$. Consequently, we have chosen θ (θ_1 or θ_2) to be the value which determines the speed since as we increase θ, which represents the rising time, we increase the time at which the maximum occurs. In other words, we decrease the speed of the signal variation.

2.2. TESTING OF THE PARAMETER SWEEPING SPEED

To test the effect of speed, we compute the ratio R numerically using

$$R = \frac{N_u}{N_t} \tag{10}$$

It represents the ratio between trajectories or cells which go to the top (or to the upper branch) of the bifurcation diagram, and trajectories that go to both upper and lower branches during simulation. Obviously, $R = 1$ if all cells choose the upper branch in the decision of their fate, $R = 0.5$ if the proportions of cells between two branches are equal, and $R = 0$ if all cells prefer the lower branch.

Heun's method is used for solving the differential equations. In simulation of stochastic differential equations we have used Matlab, and all bifurcation diagrams and nullclines were generated in XPPAUT. $\delta(t)$ is an integration step size.

3. RESULTS

3.1. GATA1-PU.1 GENETIC SWITCH WITHOUT EXTERNAL SIGNALS

This switch (**Figure 1A**) represents a paradigm for gene regulatory networks that govern the differentiation (2). It consists of two transcription factors GATA1 and PU.1 with self-stimulation and cross-inhibition. GATA1 is a master regulator of the erythroid lineage, and PU.1 is a master regulator of the myeloid lineage, and the two lineages arise from a common myeloid progenitor cell (1, 15).

3.1.1. Bifurcation analysis for symmetric scenario

In the symmetric scenario, the parameters of the model are changed symmetrically with respect to X_1 and X_2. Hence, the rates of self-activation, cross-inhibition, deactivation, and thresholds are equal for both transcription factors (see Materials and Methods). Then, this scenario is divided into two parts depending on the kind of bifurcation which results in during a change of the parameters.

- *Supercritical pitchfork bifurcation*: This type of bifurcation can occur when b is increased from 0.5 to 1 (**Figure 4A**), or when r is decreased from 1.8 to 1.2 (**Figure 5C**). In this kind of bifurcation, a transition occurs from monostability to bistability. The monostable state represents progenitor cell in undifferentiated state and has the ability to differentiate into two different fates. At this state, both transcription factors in the network are produced at approximately equal levels as it can be seen from the intersection point of nullclines in (**Figure 4B**). At the differentiation process, the progenitor cell is destabilized and two new attractors appear with equal basins of attraction (**Figure 4C**).
- *Subcritical pitchfork bifurcation*: This type of bifurcation occurs for many parameter changes. It can happen when k is changed from 1 to 1.5 (**Figure 5A**), when a is decreased from 1 to 0.5 (**Figure 5B**), when b is increased from 0.3 to 0.4 (**Figure 5D**), and when r is increased from 0.5 to 1 (**Figure 5C**). In this kind of bifurcation, a transition occurs from tristability to bistability (**Figures 5E,F**). In this situation, the progenitor cell (metastable state) coexists with the two fates, and the two transcription factors are expressed at equal or low levels. At the bifurcation process, it becomes unstable and makes discontinuous transition to either erythroid or myeloid fates with equal basins of attraction.

3.1.2. Bifurcation analysis for asymmetric scenario

Here, the parameters of the model are changed asymmetrically with respect to X_1 and X_2. For example, we can increase one of the parameters and keep the other constant, or decrease one of the parameters and keep the other constant, or both. This asymmetric change will cause symmetry-breaking in the bifurcation diagrams and makes one of the attractors more favorable than the other. Similarly, we note two types of bifurcation:

- *Supercritical pitchfork bifurcation*: In order to get this kind of bifurcation with symmetry-breaking, we increase a_1 and k_2 (see Materials and Methods for their definitions) and as a result, X_1 is increased. Then, the bifurcation occurs when b is changed from 0.6 to 1 (**Figures 6A–C**). Now, the uncommitted progenitor cell represented by monostability is not in the middle but at the point where X_1 is higher. After bifurcation, the erythroid fate becomes dominant since it has a larger basin of attraction to the right of the separatrix (**Figure 6C**).

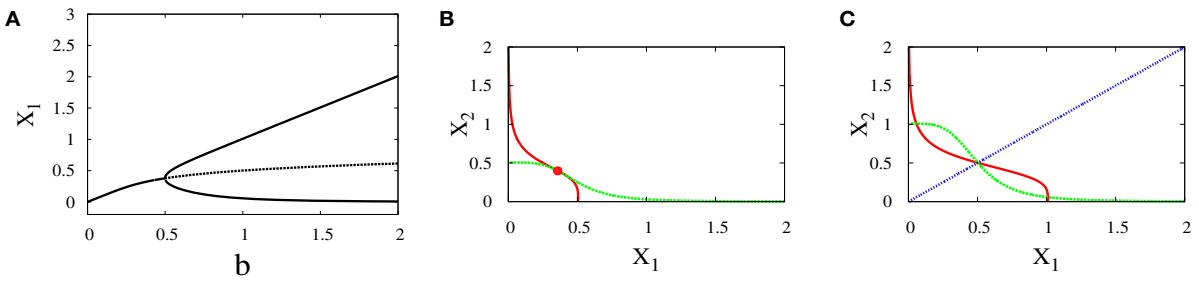


FIGURE 4 | Supercritical pitchfork bifurcation diagrams for symmetric scenario. Bifurcation diagram **(A)** and nullclines at the beginning **(B)** and end **(C)** of the bifurcation. For all diagrams, $n=4$, $r=0.5$, $a_1=a_2=0.01$, $k_1=k_2=1$. The solid lines indicate stability, while dashed lines indicate unstable branches.

- *Subcritical pitchfork bifurcation*: The bifurcation occurs when b is varied from 0.3 to 0.4. This gives imperfect subcritical pitchfork bifurcation (**Figure 7A**). The change in system behavior from tristability to bistability is depicted in (**Figures 7B,C**). At the progenitor cell, both transcription factors have low levels but the progenitor cell is not exactly in the middle. After bifurcation, one of the attractors corresponding to erythroid lineage becomes dominant as a result of increasing self-activation of GATA1.

3.1.3. Trajectories and the effect of parameter sweeping speed

To investigate the effect of the different speeds of the parameter sweeping we concentrate on the asymmetric supercritical pitchfork bifurcation, and similar results can be seen in the other kind of bifurcation. The graphical solutions of X_1 and X_2 after solving the differential equations (see equations (1) and (2) in Materials and Methods) are shown in (**Figure 8A**). As the time increases, the values of X_1 increase and the values of X_2 decrease. In fact, for small values of noise, this is the expected behavior from the dominance of the erythroid attractor.

To examine the effect of the speed with which the system crosses the critical region, we vary b linearly with time according to $b(t)=\alpha t$, where α is the slope, and compute the ratio R (see Materials and Methods). The result is shown in (**Figure 8B**). For low speeds, the ratio R is high which means that most of the cells choose the erythroid lineage due to the produced asymmetry, and this lineage leads to and include red blood cells. On the other hand, as we increase the speed, this ratio tends to zero. Two conclusions follow from this behavior. First, large speeds reduce the effect of asymmetry gradually and convert the favorable attractor completely when the ratio tends to zero. Second, $R=0$ means that the myeloid fate becomes more favorable by cells. The myeloid fate leads to the immune cells of the immune system (16).

3.2. GATA1-PU.1 GENETIC SWITCH UNDER EXTERNAL SIGNALING

To elucidate the effect of external signals on the dynamics of the switch, we consider external signals acting upon the switch (**Figure 1B**), see also equations (3) and (4). The external signals enhance the activation of X_1 and X_2. **Figure 9** highlights the bifurcation in the parameter space (S_1, S_2) for the chosen set of parameters. The borders separate between the regions of monostability and the region of bistability, and this indicates to the existence of supercritical pitchfork bifurcation under the two following scenarios.

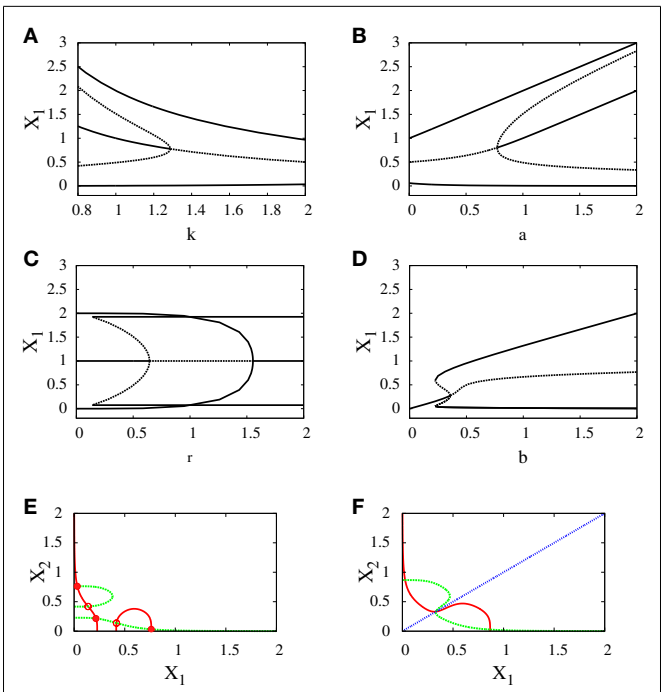


FIGURE 5 | Subcritical pitchfork bifurcation diagrams for symmetric scenario. Bifurcation diagrams **(A–D)** and nullclines at the beginning **(E)** and end **(F)** of the bifurcation diagram **(D)**. For all $n=4$. For **(A)** $a=1$, $b=1$, $r=0.5$, **(B)** $b=1$, $k=1$, $r=0.5$, **(C)** $a=1$, $b=1$, $k=1$, **(D–F)** $a_1=a_2=1$, $k_1=k_2=1.5$. In **(C)** there is also supercritical pitchfork bifurcation.

3.2.1. Bifurcation analysis for symmetric scenario

Under this scenario, both signals S_1 and S_2 are equal. The nullclines in (**Figures 10A,B**) exhibit the bifurcation from monostability to bistability. This symmetry will give us near-symmetric bifurcation diagram (**Figure 10C**) with progenitor cell located in the middle and have equal probabilities to choose between erythroid (upper attractor) and myeloid (lower attractor) fates.

3.2.2. Bifurcation analysis for asymmetric scenario

In contrary to the above scenario, now the signals have different parameters. As a result, the monostable state (**Figure 10D**) is at the point where X_1 is higher since S_1 which acts on X_1 is larger. After bifurcation, the attractor at which X_1 is high has a larger

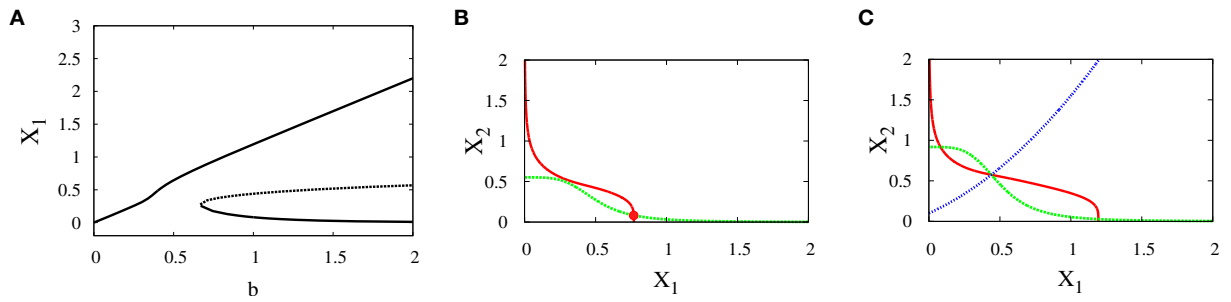


FIGURE 6 | Asymmetric scenario. Supercritical pitchfork bifurcation diagrams. Bifurcation diagram **(A)** and nullclines at the beginning **(B)** and end **(C)** of the bifurcation. The parameters are $n=4$, $r=0.5$, $a_1=0.2$, $a_2=0.01$, $k_1=1$, $k_2=1.1$.

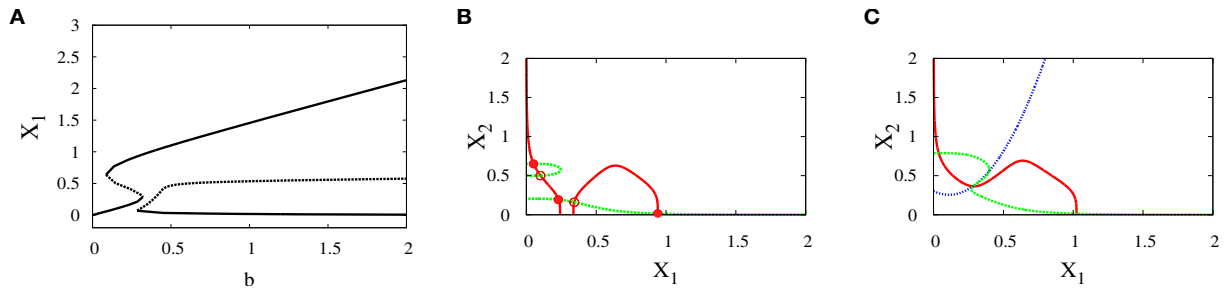


FIGURE 7 | Asymmetric scenario. Subcritical pitchfork bifurcation diagrams. Bifurcation diagram **(A)** and nullclines at the beginning **(B)** and end **(C)** of the bifurcation. The parameters are $n=4$, $r=0.5$, $a_1=1.2$, $a_2=1$, $k_1=1.5$, $k_2=1.6$.

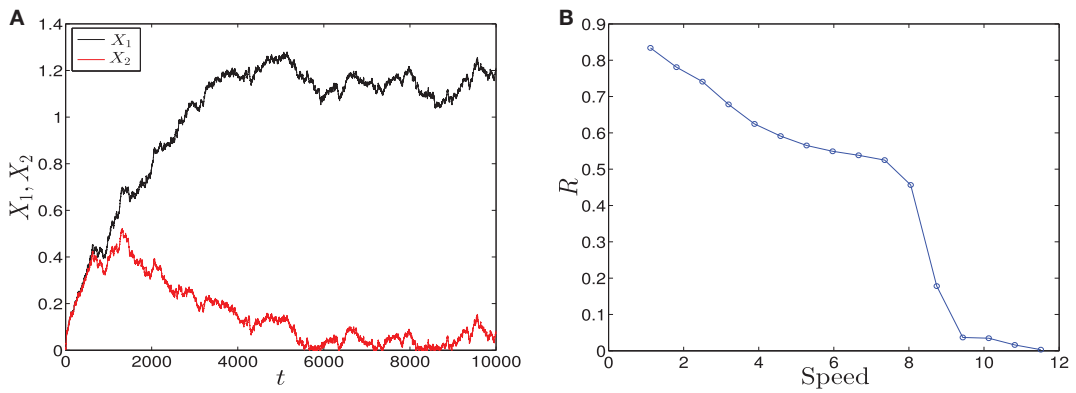


FIGURE 8 | Trajectories and parameter sweeping speed. **(A)** Time evolution of X_1 and X_2 in the asymmetric supercritical pitchfork bifurcation. **(B)** The effect of increasing the speed of crossing the critical region on the distribution of trajectories in the attractors for 10000 iterations. As the speed is increased, the ratio R changes from 1 to 0. Hence, increasing the speed causes a large switch from the favorable attractor to the other one. Parameters are $a_1=0.2$, $a_2=0.01$, $k_1=1$, $k_2=1.1$, $n=4$, $r=0.5$. Also, in **(A)** $\sigma^2=0.01$, **(B)** $\sigma^2=0.5$.

basin of attraction (**Figure 10E**). We can note in (**Figure 10F**) how this asymmetry produces symmetry-breaking in the bifurcation diagram and so the decision of the cell will be biased.

3.2.3. Trajectories and speed-dependent cellular decision making

To study how signal asymmetry, noise, and decision making will result in the dependence of the parameter sweeping speed we have considered with two kinds of signals (See Materials and Methods):

- *Linear signals*: The signals are shown in (**Figure 2A**). The asymmetry between the two signals is transient and the symmetry is retained after some time (**Figure 2B**). The behavior of trajectories of X_1 and X_2 under the influence of this form of signals

is shown in (**Figure 11A**). As the time increases, the values of X_1 increase and the values of X_2 decrease. Hence, trajectories of X_1 and X_2 choose the attractor at which X_1 is higher since S_1 is faster. Next, to test the effect of increasing the speed on choosing the attractors (stable steady states) of genetic switch in the presence of noise and transient asymmetry A, we vary T_1 in $S_1(t) = \frac{S_{max}}{T_1} t$ with constant values of A and S_{max}, and T_2 will be changed according to the formula $T_2 = \frac{S_{max}}{S_{max}-A} T_1$ (7). With increasing the speed (**Figure 11B**), the ratio R tends to 0.5. Thus, increasing the speed increases the symmetry between erythroid and myeloid lineages and reduce the effect of asymmetry which is produced by external signals.

- *Adaptation form of signals*: The signals take the non-linear form shown in (**Figure 3**) and as for the linear form, S_1 is faster than S_2. The trajectories in this form behave almost like the first form (**Figure 11C**). To study the effect of the speed, we vary θ_1 and θ_2 such that θ_1 is smaller than θ_2. Then, we compute the ratio R and the result is depicted in (**Figure 11D**). It shows ratio tending to 0.5 as θ is increased. But increasing θ decreases the speed, so, surprisingly, now we regain the symmetry in the switch by decreasing the speed of external signals.

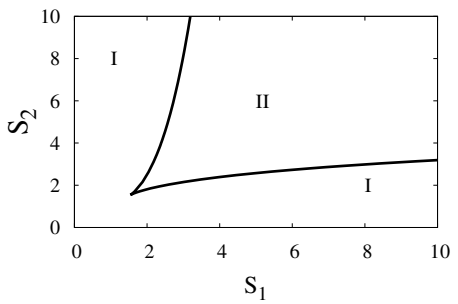


FIGURE 9 | Two-parameter bifurcation diagram. The bifurcation in the parameter space (S_1, S_2), where S_1 and S_2 are external signals in the genetic switch. The borders separate between the regions of monostability *I* and the region of bistability *II*. Parameters are $a_1 = a_2 = 0.05$, $b_1 = b_2 = 0.45$, $r = 0.5$, $k_1 = k_2 = 1$.

4. DISCUSSION

We have shown the importance of parameter sweeping speed when the gene regulatory circuit of immune cell differentiation is exposed to external factors that cause symmetry-breaking and make one of the attractors or fates more favorable than the other. In our study, symmetry-breaking is caused by three factors. The first factor is the asymmetric change of parameters which gives ratio tends to zero as the speed is increased (**Figure 8B**). This means we get large conversion from the favorite attractor, the erythroid lineage, to the myeloid lineage. The importance of this effect may appear in cases where the person has a problem with immunity due to the decrease in the production of immune cells, so even when there is a bias in the cell and this bias has the effect of choosing the attractor where GATA1 is upregulated, the cell can be forced to

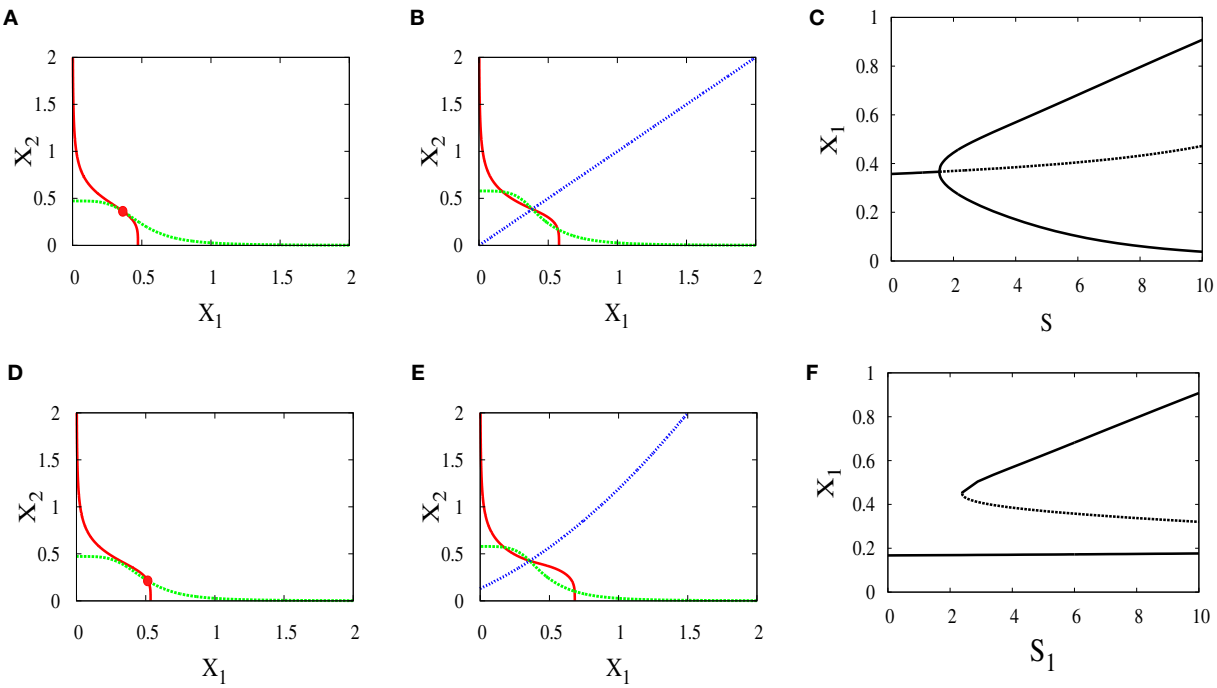


FIGURE 10 | Nullclines and bifurcation diagrams with symmetric and asymmetric external signals. For all figures, $a_1 = a_2 = 0.05$, $b_1 = b_2 = 0.45$, $r = 0.5$, $k_1 = k_2 = 1$, $n = 4$. **(A)** Nullclines for $S_1 = S_2 = 1$ show one stable steady state, **(B)** nullclines for $S_1 = S_2 = 4$ show bistability, **(C)** near-symmetric supercritical pitchfork bifurcation, **(D)** nullclines for $S_1 = 3$, $S_2 = 1$ show one stable steady state shifted to the right, **(E)** nullclines for $S_1 = 6$, $S_2 = 4$ show bistability with larger basin of attraction to the right of the separatrix (almost diagonal line), **(F)** imperfect supercritical pitchfork bifurcation due to the asymmetry between the external signals.

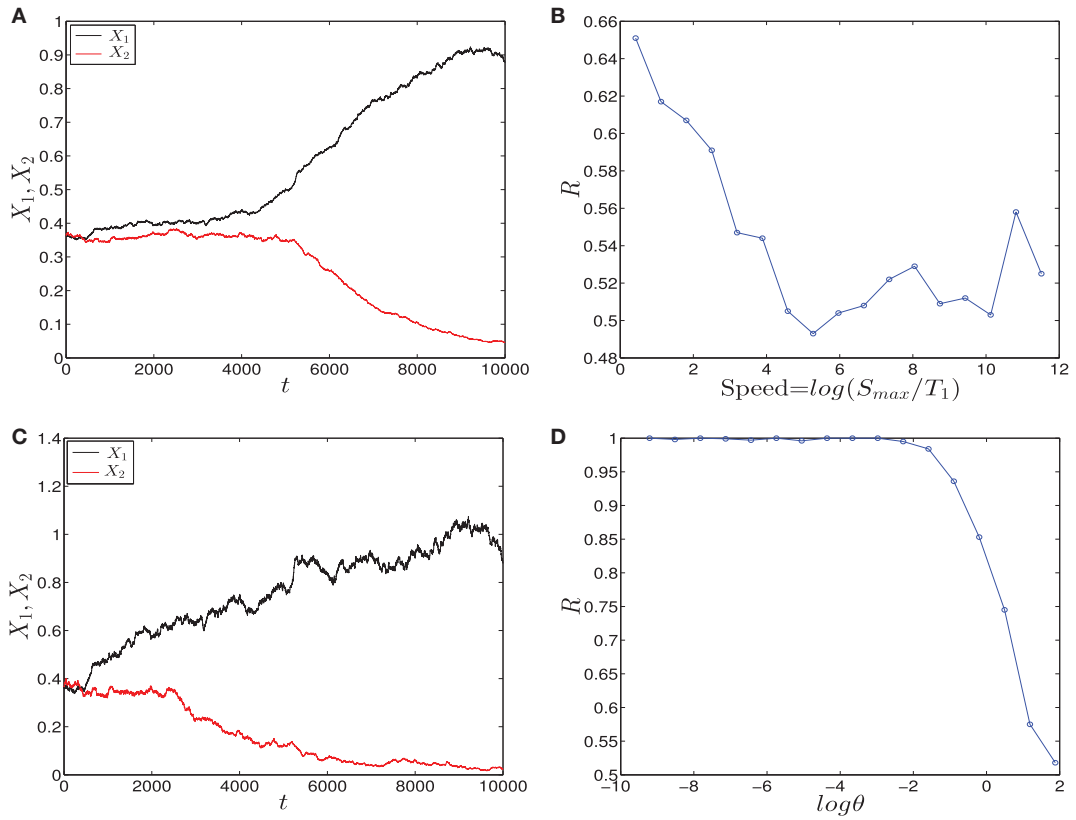


FIGURE 11 | Trajectories and effect of the external signaling speed. In **(A,C)** the time evolution of X_1 and X_2 under the effect of linear and adaptation form of signals is shown, respectively. The values of X_1 increase and the values of X_2 decrease because S_1 is chosen to be faster than S_2. Hence, trajectories choose the attractor which has a larger value of X_1. **(B)** The effect of increasing the speed of linear form of signals for 1000 iterations. As the speed is increased, the ratio R tends to 0.5. Thus, increasing the speed increases the symmetry in the switch. **(D)** The effect of speed with the adaptation form of signals. Decreasing the speed gives ratio R tending to 0.5. Surprisingly, now decreasing the speed increases the symmetry in the switch. Parameters in **(A)**, **(B)** are $A = 2.5$, $S_{max} = 10$, **(C,D)** $h_1 = h_2 = 10$, $v = 10$, and for all we have $a_1 = a_2 = 0.05$, $b_1 = b_2 = 0.45$, $r = 0.5$, $k_1 = k_2 = 1$.

choose the attractor where PU.1 is upregulated by increasing the speed of crossing the critical region and so enhancing the production of immune cells. The second factor is linear form of signals (**Figure 2A**) and in this case we get ratio tends to 0.5 with increasing the speed (**Figure 11B**). This result may be important in situations that need symmetry between erythroid and myeloid cells, or when decreasing the probability of choosing the erythroid lineage is required. The third factor is represented by non-linear form of signals, i.e., signals describing biochemical adaptation (**Figure 3**). Here, decreasing the speed blinds the asymmetry and produce symmetry between the two lineages (**Figure 11D**). Taken together, the external signals, its shape, and its speed may have critical effects on choosing the attractors and affect the cell-fate determination.

Notably, we followed the model of Huang et al. (2) to study the differentiation into erythroid and myeloid fates. On the other hand, there is a scheme in Ref. (1, 17) that gives additional kinds of cells or lineages under each transcription factor. In this scheme, GATA1 is responsible for differentiation into erythroid or megakaryocyte cells, and PU.1 leads to either lymphoid lineage (B and T cells) which gives the *Adaptive Immune Cells*, or myeloid lineage (macrophages and granulocytes) that produces the *Innate Immune Cells*. So for this scheme, the importance of parameter sweeping speed is increased as the fate corresponding to high concentration of PU.1 is able to produce the different types of immune cells.

Of particular interest and agreement with our conclusions about the importance of external signals, Heuser et al. (6) have showed the crucial role of external signals in MN1 leukemia. They have investigated the requirement of FLT3 and c-Kit signals for MN1 leukemia. Overexpression of MN1 induces myeloid leukemia and blocks erythroid differentiation. FLT3 and c-Kit signaling direct MN1-expressing cells toward the myeloid lineage, so disruption of these signals may prevent leukemia. Interestingly, the disruption of these external signals doesn't delay the disease latency but induces a switch from myeloid to erythroid lineage. Thus, the external signals can alter leukemia stem cell differentiation fates.

Many models have focused on the role of external signals in the differentiation process (3, 5) but they haven't given any attention to the shape of signals or to the speed of these signals. Additionally, many works have made their studies limited to the symmetric scenario for the sake of simplicity (2, 18). But in this paper, we have studied the asymmetric scenarios and investigated the effect of

external signaling speed on the system's dynamics. As a prospect, it would be specially interesting to study the effect of speed on more complicated models and including other factors that may have a role in the differentiation process of hematopoietic stem cells, which can lead to better understanding of the immune system. Furthermore, an experimental evidence is needed to support the predictions from the mathematical models.

ACKNOWLEDGMENTS

This project was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University (KAU), Jeddah, under grant No. (20/34/Gr). Alexey Zaikin acknowledges support from CR-UK and Eve Appeal (PROMISE-2016).

REFERENCES

1. Laslo P, Pongubala JMR, Lancki DW, Singh H. Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin Immunol* (2008) **20**:228–35. doi:10.1016/j.smim.2008.08.003
2. Huang S, Guo Y, May G, Enver T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol* (2007) **305**:695–713. doi:10.1016/j.ydbio.2007.02.036
3. Chickarmane V, Enver T, Peterson C. Computational modeling of the hematopoietic erythroid-myeloid switch reveals insight into cooperativity, priming, and irreversibility. *PLoS Comput Biol* (2009) **5**(1):e1000268. doi:10.1371/journal.pcbi.1000268
4. Palani S, Sarkar CA. Positive receptor feedback during lineage commitment can generate ultrasensitivity to ligand and confer robustness to a bistable switch. *Biophys J* (2008) **95**:1575–89. doi:10.1529/biophysj.107.120600
5. Palani S, Sarkar CA. Integrating extrinsic and intrinsic cues into a minimal model of lineage commitment for hematopoietic progenitors. *PLoS Comput Biol* (2009) **5**(9):e1000518. doi:10.1371/journal.pcbi.1000518
6. Heuser M, Park G, Moon Y, Berg T, Xiang P, Kuchenbauer F. Extrinsic signals determine myeloid-erythroid lineage switch in MN1 leukemia. *Exp Hematol* (2010) **38**:174–9. doi:10.1016/j.exphem.2010.01.003
7. Nene NR, Garca-Ojalvo J, Zaikin A. Speed-dependent cellular decision making in nonequilibrium genetic circuits. *PLoS One* (2012) **7**(3):e32779. doi:10.1371/journal.pone.0032779
8. Nene NR, Zaikin A. Interplay between path and speed in decision making by high-dimensional stochastic gene regulatory networks. *PLoS One* (2012) **7**(7):e40085. doi:10.1371/journal.pone.0040085
9. Nene NR, Zaikin A. Decision making in noisy bistable systems with time-dependent asymmetry. *Phys Rev E Stat Nonlin Soft Matter Phys* (2013) **87**:012715. doi:10.1103/PhysRevE.87.012715
10. Ashwin P, Wieczorek S, Vitolo R, Cox P. Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system. *Philos Trans A Math Phys Eng Sci* (2012) **370**:1166–84. doi:10.1098/rsta.2011.0306
11. Young JW, Locke JCW, Elowitz MB. Rate of environmental change determines stress response specificity. *Proc Natl Acad Sci U S A* (2013) **110**:4140–5. doi:10.1073/pnas.1213060110
12. Weiss JN. The Hill equation revisited: uses and misuses. *FASEB J* (1997) **11**:835–41.
13. Gillespie DT. The chemical Langevin equation. *J Chem Phys* (2000) **113**:297–306. doi:10.1063/1.481811
14. Ma W, Trusina A, El-Samad H, Lim WA, Tang C. Defining network topologies that can achieve biochemical adaptation. *Cell* (2009) **138**:760–73. doi:10.1016/j.cell.2009.06.013
15. Cinquin O, Demongeot J. High-dimensional switches and the modelling of cellular differentiation. *J Theor Biol* (2005) **233**:391–411. doi:10.1016/j.jtbi.2004.10.027
16. Kawamoto H, Katsura Y. A new paradigm for hematopoietic cell lineages: revision of the classical concept of the myeloid-lymphoid dichotomy. *Cell* (2009) **30**:193–200. doi:10.1016/j.it.2009.03.001
17. Adolfsson J, Mansson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, et al. Identification of $Flt3^+$ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment. *Cell* (2005) **121**:295–306. doi:10.1016/j.cell.2005.02.013
18. Foster DV, Foster JG, Huang S, Kauffman SA. A model of sequential branching in hierarchical cell fate determination. *J Theor Biol* (2009) **260**:589–97. doi:10.1016/j.jtbi.2009.07.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 August 2013; paper pending published: 12 October 2013; accepted: 20 November 2013; published online: 05 December 2013.
*Citation: Alagha A and Zaikin A (2013) Asymmetry in erythroid-myeloid differentiation switch and the role of timing in a binary cell-fate decision. Front. Immunol. **4**:426. doi: 10.3389/fimmu.2013.00426*
This article was submitted to T Cell Biology, a section of the journal Frontiers in Immunology.
Copyright © 2013 Alagha and Zaikin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.