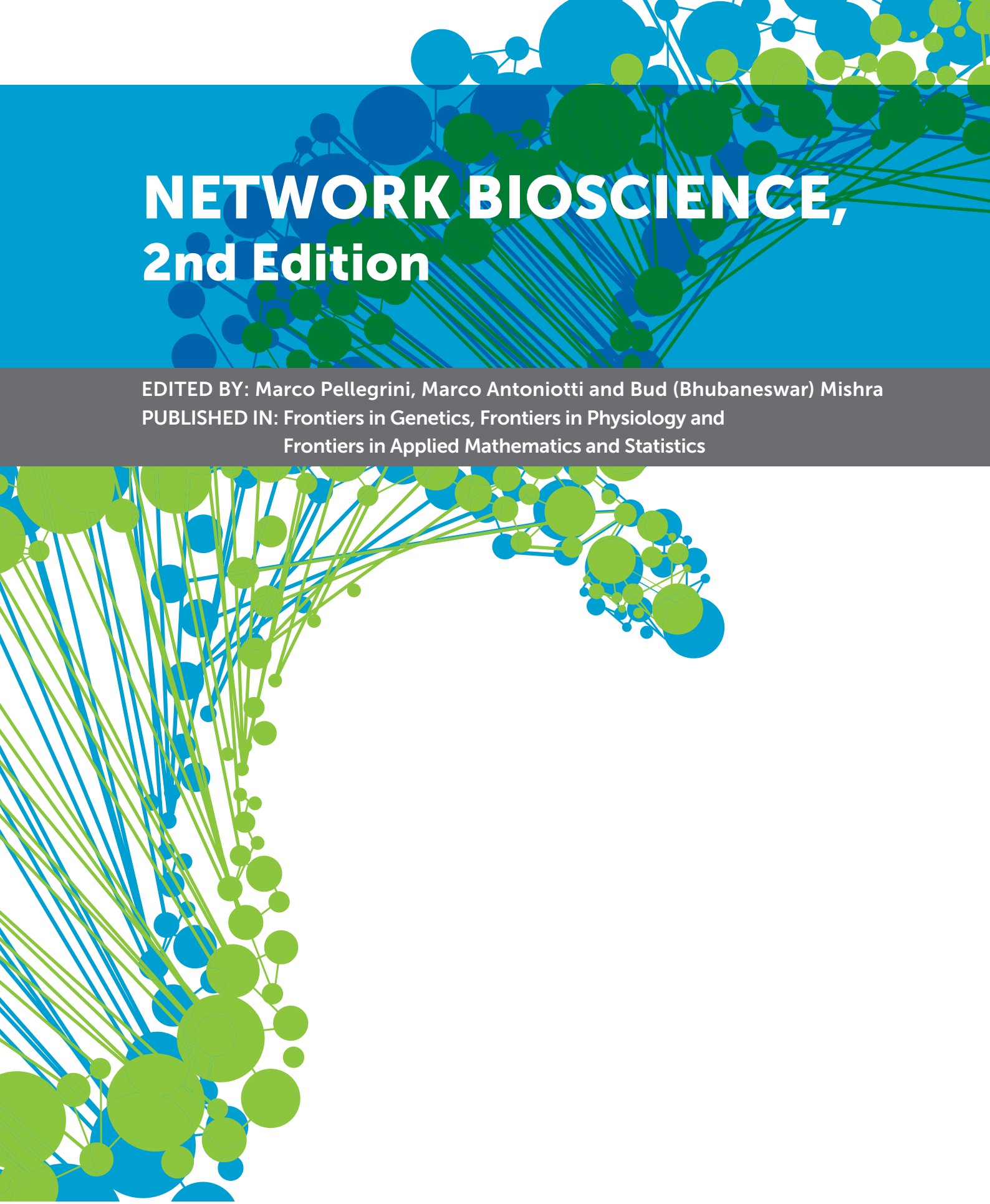


NETWORK BIOSCIENCE, 2nd Edition

EDITED BY: Marco Pellegrini, Marco Antoniotti and Bud (Bhubaneswar) Mishra

PUBLISHED IN: Frontiers in Genetics, Frontiers in Physiology and
Frontiers in Applied Mathematics and Statistics

frontiers Research Topics

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88963-650-1
DOI 10.3389/978-2-88963-650-1

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

NETWORK BIOSCIENCE, 2nd Edition

Topic Editors:
Marco Pellegrini, Italian National Research Council, Italy
Marco Antoniotti, University of Milano Bicocca, Italy
Bud (Bhubaneswar) Mishra, New York University, United States

Publisher's note: In this 2nd edition, the following article has been updated: Yepiskoposyan H, Talikka M, Vavassori S, Martin F, Sewer A, Gubian S, Luettich K, Peitsch MC and Hoeng J (2019) Construction of a Suite of Computable Biological Network Models Focused on Mucociliary Clearance in the Respiratory Tract. Front. Genet. 10:87. doi: 10.3389/fgene.2019.00087

Citation: Pellegrini, M., Antoniotti, M., Mishra, B., eds. (2020). Network Bioscience, 2nd Edition. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-650-1

Table of Contents

05 ***Editorial: Network Bioscience***
Marco Antoniotti, Bud Mishra and Marco Pellegrini

07 ***On the Origin of Biomolecular Networks***
Heeralal Janwa, Steven E. Massey, Julian Velev and Bud Mishra

25 ***Metabolic Games***
Taneli Pusa, Martin Wannagat and Marie-France Sagot

38 ***Network Medicine in the Age of Biomedical Big Data***
Abhijeet R. Sonawane, Scott T. Weiss, Kimberly Glass and Amitabh Sharma

54 ***Comparative Analysis of Normalization Methods for Network Propagation***
Hadas Biran, Martin Kupiec and Roded Sharan

62 ***To Embed or Not: Network Embedding as a Paradigm in Computational Biology***
Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg and Roded Sharan

73 ***Comprehensive Review of Models and Methods for Inferences in Bio-Chemical Reaction Networks***
Pavel Loskot, Komlan Atitey and Lyudmila Mihaylova

102 ***Construction of a Suite of Computable Biological Network Models Focused on Mucociliary Clearance in the Respiratory Tract***
Hasmik Yepiskoposyan, Marja Talikka, Stefano Vavassori, Florian Martin, Alain Sewer, Sylvain Gubian, Karsta Luettich, Manuel Claude Peitsch and Julia Hoeng

114 ***Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome***
Ryan A. Miller, Friederike Ehrhart, Lars M. T. Eijssen, Denise N. Slenter, Leopold M. G. Curfs, Chris T. Evelo, Egon L. Willighagen and Martina Kutmon

124 ***A Comprehensive Survey of Tools and Software for Active Subnetwork Identification***
Hung Nguyen, Sangam Shrestha, Duc Tran, Adib Shafi, Sorin Draghici and Tin Nguyen

139 ***NoMAS: A Computational Approach to Find Mutated Subnetworks Associated With Survival in Genome-Wide Cancer Studies***
Federico Altieri, Tommy V. Hansen and Fabio Vandin

151 ***pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks***
Ege Ulgen, Ozan Ozisik and Osman Ugur Sezerman

184 ***Adapting Community Detection Algorithms for Disease Module Identification in Heterogeneous Biological Networks***
Beethika Tripathi, Srinivasan Parthasarathy, Himanshu Sinha, Karthik Raman and Balaraman Ravindran

201 ***QS-Net: Reconstructing Phylogenetic Networks Based on Quartet and Sextet***
Ming Tan, Haixia Long, Bo Liao, Zhi Cao, Dawei Yuan, Geng Tian, Jujuan Zhuang and Jialiang Yang

210 ***A Computational Pipeline for the Extraction of Actionable Biological Information From NGS-Phage Display Experiments***
Antonios Vekris, Eleftherios Pilalis, Aristotelis Chatziioannou and Klaus G. Petry

221 ***ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next Generation Sequencing Data***
Theodoros Koutsandreas, Efthymios Ladoukakis, Eleftherios Pilalis, Dimitra Zarafeta, Fragiskos N. Kolisis, Georgios Skretas and Aristotelis A. Chatziioannou

236 ***A Multi-Cohort and Multi-Omics Meta-Analysis Framework to Identify Network-Based Gene Signatures***
Adib Shafi, Tin Nguyen, Azam Peyvandipour, Hung Nguyen and Sorin Draghici

252 ***Multi-Phenotype Association Decomposition: Unraveling Complex Gene-Phenotype Relationships***
Deborah Weighill, Piet Jones, Carissa Bleker, Priya Ranjan, Manesh Shah, Nan Zhao, Madhavi Martin, Stephen DiFazio, David Macaya-Sanz, Jeremy Schmutz, Avinash Sreedasyam, Timothy Tschaplinski, Gerald Tuskan and Daniel Jacobson

frontiers in Genetics

EDITORIAL
published: 20 November 2019
doi: 10.3389/fgene.2019.01160

Editorial: Network Bioscience

Marco Antoniotti[1], *Bud Mishra*[2] *and Marco Pellegrini*[3]*

[1] *Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy,* [2] *Courant Institute of Mathematical Sciences, New York University, New York, NY, United States,* [3] *Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Pisa, Italy*

Keywords: systems biology, network science, network biology, cancer networks, hypothesis generation and verification, computational biology

Editorial on the Research Topic

Network Bioscience

NETWORKS IN MANY GUISES FOR BIOSCIENCE

In the last decade, the very nature of biological research has changed as large-scale data arrive at torrential force and it has ushered in a new era of *Bioscience*; but also this high dimensional big data is being used to support inference of various types and multiplicities of hypotheses about the extant relationships among the "variables" being measured.

The typical current example in the biomedical field is *sequencing data* (in various forms: *DNA sequencing*, *RNA sequencing*, *ATAC using sequencing*, etc.). Another kind of data currently collected is *proteomic* data, often with the goal of producing *protein-protein interaction networks* (PPI networks). Yet another is data about the *metabolome* of a biological system. Moreover recently, also phenotypic data, data on diseases, symptoms, patients, etc., are being collected at nation-wide level thus giving us another source of highly related (causal) "big data."

From these kinds of data, biologists and bioinformaticians, can make many inferences, and, more often than not, such inferences now reuse several notions, theories, and tools from the field of *network science*. Network science has accelerated a deep and successful trend in research that influences a range of disciplines like mathematics, graph theory, physics, statistics, data science, and computer science (just to name a few), and adapts the relevant techniques and insights to address relevant but disparate social, biological, technological questions.

Most of the data kinds just mentioned naturally lend themselves to a *network* analysis. The network model is a key viewpoint leading to the uncovering of mesoscale phenomena, thus providing an essential bridge between the observable phenotypes and *omics* underlying mechanisms. Moreover, network analysis is a powerful *hypothesis generation* tool guiding the scientific cycle of *data gathering*, *data interpretation*, *hypothesis generation,* and *hypothesis testing*.

The papers contained in the present research topic—*Network Bioscience*—are examples of how network and graph analysis can be used to elucidate various aspect of biological systems from metabolic regimes, to phenotype-genotype linking, to relationships assessment among diverse omics data for therapy design, to functional submodule identification in a gene network for cancer studies.

PAPERS PRESENTATION

The papers collected in this research topic are roughly grouped as follows:

• "Foundational" papers,

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Marco Pellegrini
marco.pellegrini@iit.cnr.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *03 October 2019*
Accepted: *23 October 2019*
Published: *20 November 2019*

Citation:
Antoniotti M, Mishra B and Pellegrini M (2019) Editorial: Network Bioscience. Front. Genet. 10:1160. doi: 10.3389/fgene.2019.01160

- Analysis of particular biomedical problems,
- Tool presentations.

Several contributions tackle foundational aspects of network bioscience, relative to their origin, evolution, underlying philosophy, mathematical modelling, as well as connections to network medicine on one hand and dynamics of bio-chemical reactions on the other hand.

Janwa et al. explore the role of information asymmetry in the genesis and evolution of pairwise biomolecular interactions leading to the formation of extensive and complex networks of biomolecular interactions. Pusa et al. review a network-based evolutionary game-theoretic view of emerging phenotypes and its use in the context of metabolic modeling. Sonawane et al. connect the emerging field of network medicine with the opportunity of collecting big biomedical data identifying three different network archetypes according to different underlying philosophies.

Biran et al. and Nelson et al. discuss mathematical aspects of bio-networks science relative to the benefits of propagation of information in bio-networks and the benefits of embedding bio-networks into low-dimensional Euclidean spaces both for visualization and for tasks such as network de-noising, modularization, and function prediction.

Loskot et al. survey recent advances in the broad area of bio-chemical reaction networks, which constitute a crucial model for elucidating non-linear dynamics of bio-chemical processes.

Two papers report on the application of network-based models to unravel complex physiological and pathological processes, namely the molecular mechanisms causing mucociliary clearance in the human respiratory tract (Yepiskoposyan et al.) and the role of active regulatory sub-networks characterizing a genetic brain disorder: Rett Syndrome (Miller et al.).

Active subnetwork/module identification is a key step in the process of discovering differences between cases and controls (e.g., pathological and healthy states) that fully exploit the rich structure of the bio-network models, and play a key role parallel to that of DGE (differential gene expression) in comparative genomic expression analysis. Nguyen et al. contributed a review of 22 state-of-the-art integrative tools and algorithms for such problem, including a discussion of outstanding challenges and open problems. Two new original methods: *NoMAS* by Altieri et al. and *PathFindR* Ulgen et al. push forward the state of the art on active subnetwork/module identification. Tripathi et al. discuss important issues relative to benchmarking of active subnetwork/module identification methods, and to the adaptation of existing general-purpose community detection algorithms for this task.

Converting raw-data into a suitable network model is a non-trivial task, a source of very important and challenging problems. Here we have a few such examples. Tan et al. describe *QS-Net*, an accurate methodology for building phylogenetic networks from basic sequencing data. Vekris et al. develop analytical tools and strategies for de-noising phage display data, employing graph-theoretic methods. Koutsandreas et al. report on the new pipeline *ANASTASIA* for metagenomic analysis of environmental samples, which is a challenging source of data. Shafi et al. focus on the challenge of defining multi-cohort and multi-omics meta-analysis framework that overcomes limitations of less integrative approaches in order to identify robust molecular subnetworks that capture the key dynamic nature of a given biological condition. Weighill et al. unravel the multi-phenotype signatures of genes on a genome-wide network built from SNP-phenotype association data, thus improving the interpretation of large GWAS datasets and aiding in future synthetic biology efforts designed to optimize phenotypes of interest.

AUTHOR CONTRIBUTIONS

The editors all contributed equally to the research topic assembly and editing and to this editorial.

FUNDING

This work was partially supported by the CRUK/AIRC/FC-AECC Accelerator Award #22790, *Single Cell Cancer Evolution in the Clinic*, by the FAQC 2018 *Competitive Funding Program* of the Università degli Studi di Milano-Bicocca, Milan, Italy; a National Cancer Institute Physical Sciences-Oncology Center Grant U54 CA193313-01; and by the European COST Action #CA15110 *CHARME: Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research*. This work has been partially supported within project PRIN 201534HNXC "The role of tandem repeats in neurodegenerative diseases: a genomic and proteomic approach" funded by the Italian Ministry of Education and University (MIUR) and by the National Cancer Institute Physical Sciences-Oncology Center Grant U54 CA193313-01.

Conflict of Interest: 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.

Copyright © 2019 Antoniotti, Mishra and Pellegrini. This is an openaccess 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) and the copyright owner(s) 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 Genetics

SYSTEMATIC REVIEW
published: 10 April 2019
doi: 10.3389/fgene.2019.00240

On the Origin of Biomolecular Networks

Heeralal Janwa[1], *Steven E. Massey*[2], *Julian Velev*[3] *and Bud Mishra*[4]*

[1] Department of Mathematics, University of Puerto Rico, San Juan, PR, United States, [2] Department of Biology, University of Puerto Rico, San Juan, PR, United States, [3] Department of Physics, University of Puerto Rico, San Juan, PR, United States, [4] Departments of Computer Science, Mathematics and Cell Biology, Courant Institute and NYU School of Medicine, New York University, New York City, NY, United States

Biomolecular networks have already found great utility in characterizing complex biological systems arising from pairwise interactions amongst biomolecules. Here, we explore the important and hitherto neglected role of information asymmetry in the genesis and evolution of such pairwise biomolecular interactions. Information asymmetry between sender and receiver genes is identified as a key feature distinguishing early biochemical reactions from abiotic chemistry, and a driver of network topology as biomolecular systems become more complex. In this context, we review how graph theoretical approaches can be applied not only for a better understanding of various proximate (mechanistic) relations, but also, ultimate (evolutionary) structures encoded in such networks from among all types of variations they induce. Among many possible variations, we emphasize particularly the essential role of gene duplication in terms of *signaling game theory*, whereby sender and receiver gene players accrue benefit from gene duplication, leading to a preferential attachment mode of network growth. The study of the resulting dynamics suggests many mathematical/computational problems, the majority of which are intractable yet yield to efficient approximation algorithms, when studied through an algebraic graph theoretic lens. We relegate for future work the role of other possible generalizations, additionally involving horizontal gene transfer, sexual recombination, endo-symbiosis, etc., which enrich the underlying graph theory even further.

Keywords: biomolecules, regulation and communication, interaction (binary) relationship, network model, network analysis, spectral analysis

OPEN ACCESS

Edited by:
Seungchan Kim,
Prairie View A&M University, United States

Reviewed by:
Sungwon Jung,
Gachon University of Medicine and Science, South Korea
Matteo Brilli,
University of Milan, Italy

***Correspondence:**
Bud Mishra
mishra@nyu.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *04 December 2018*
Accepted: *04 March 2019*
Published: *10 April 2019*

Citation:
Janwa H, Massey SE, Velev J and Mishra B (2019) On the Origin of Biomolecular Networks. Front. Genet. 10:240. doi: 10.3389/fgene.2019.00240

1. GENESIS OF BIO-MOLECULAR INTERACTIONS

1.1. Introduction and a Road Map

A range of complex phenotypes of biomolecular systems can be inferred from macromolecular interactions, represented using combinatorial networks. Such biomolecular networks include gene (regulatory) networks (GRNs) (Thompson et al., 2015), protein-protein interaction (PPI) networks (Huang et al., 2017), protein and RNA neutral networks (Schuster et al., 1994; Govindarajan and Goldstein, 1997), metabolic networks (McCloskey et al., 2013), and meta-metabolic networks (composite metabolic networks of communities) (Yamada et al., 2011). Here, we will focus on the neglected role of information asymmetry between genes and their gene products, which is identified as a key factor distinguishing biochemistry from abiotic chemistry in early life, and which has subsequently influenced biochemical processes. Such pairwise interactions led to the establishment

of the earliest biomolecular networks, and their nature influenced subsequent network growth. We will concentrate on GRNs and PPI networks as illustrative examples, but the principles outlined are also applicable to the other types of biomolecular networks. We focus on mathematical and algorithmic techniques that by analyzing evolutionary dynamics may shed light on possible approaches to speculate on the very "origin of networks," and the challenges they pose. For simplicity, we illustrate the approaches highlighting "Evolution by Duplication" (EBD); other dynamics may be handled *mutatis mutandis*.

The paper will adhere to the following road-map, aimed at identifying and explaining several challenges for the field of "evolutionary biology of networks" by first building on a review of biological and mathematical notions and frameworks, within which the open questions are formulated. First, a brief introduction presents biomolecular networks, and biomolecular signaling games between gene players. Second, it is followed by a consideration of the role of gene duplication from the perspective of information asymmetry. Switching to a mathematical formulation, a compendium of known results in (algebraic and combinatorial) graph theory are presented, comprising a toolbox for addressing the topics raised here. Last but not least, a series of open problems are described. These open problems focus largely on the following: How to devise efficient (algebraic) algorithms that can shed important lights on *game theoretic models of the evolution of biomolecular interactions*, given that they are driven by information asymmetry (leading to duplications, complementation, pseudogenization, etc.). Some of these important mechanisms have been studied qualitatively elsewhere, albeit not mathematically rigorously.

1.2. Ohno's Evolution by Duplication

At the genetic level, the growth of a GRN (gene regulatory) or PPI (protein-protein interaction) network is driven by gene mutation, including duplication, translocation, inversion, deletion, short indels, and point mutations, of which duplication plays an outsized role, although as we incorporate other known and unknown mechanisms (e.g., non-orthologous gene displacement, HGT, sexual recombination, etc.) a more complete picture may emerge. Susumu Ohno coined the phrase "evolution by duplication" (EBD) to emphasize duplication in the evolutionary dynamic (Ohno, 1970). Consequently, we will mainly consider the process of gene duplication, but the principles outlined may be regarded as an idealization, which may be extended to other mutational processes—some yet to be discovered.

The classic view of molecular evolution is that gene families may expand and contract over evolutionary time largely due to gene duplication and deletion (Demuth et al., 2018). Here, we wish to present a more complex view, by exploring how biomolecular networks may grow, contract, or alter their topology over time, from the relative dynamic contributions and interactions of their constituent genes and gene families, and we do so through the prism of signaling game theory. Mechanistically, this evolution is driven in large part by the process of gene duplication and deletion, which lead to node and edge addition, or removal, from a biomolecular network, respectively. Since such variations in the network alter the phenotypes, over which selection operates, the evolution of networks and their features ultimately capture the essence of Darwinian evolution.

Recently, we introduced a signaling games perspective of biochemistry and molecular evolution (Massey and Mishra, 2018). There, we focused on interactions between biological macromolecules, which may be described using the framework of sender-receiver signaling games, where an expressed macromolecule such as a protein or RNA, constitutes a signal sent on behalf of a sender agent (e.g., gene). The signal comprises the three-dimensional (3D) conformation and physico-chemical properties of the macromolecule. A receiver agent (e.g., a gene product, another macromolecule) may then bind to the signal macro-molecule, which produces an action (such as an enzymatic reaction). The action produces utility for the participating agents, sender and receiver, and thereby—albeit indirectly—a change in overall fitness of the genome (in evolutionary game theory, utility and fitness are treated as analogous). When there is common interest, the utility is expected to benefit both sender and receiver and their selection, thus driving Darwinian evolution.

Replicator dynamics allow the signaling game to be couched in evolutionary terms (Taylor and Jonker, 1978). These arise from the increased replication of players with higher utility (fitness). Thus, if a gene has a strategy that results in increased utility, then it will increase in frequency in a population. For a sender gene this would entail sending a signal that results in an increase in utility, while for a receiver gene this would entail undertaking an action that likewise results in an increase in utility. As already suggested, these dynamics represent a process analogous to Darwinian (adaptive) evolution or positive selection.

Biomolecular signaling games are sustained by information asymmetry between sender and receiver, and so their interactions can be represented using directed graphs (as defined in section 2). Information asymmetry arises because the receiver is uninformed regarding the identity of the sender gene: it must rely on the signal macromolecule to determine its identity. But, this strategy may be open to deception. However, most biomolecular signaling games in the cell are between sender and receiver genes which have perfect common interest. This is so, because they are *cellularized*, chromosome replication is synchronized and so the genes replicate in concert. Such games are termed "Lewis signaling games," and rely on honest signaling from sender to the receiver, which constitutes a signaling convention (Lewis, 1969). A biomolecular signaling game is illustrated in **Figure 1**, part (1).

On occasion, situations may arise where a sender has a conflict of interest with the receiver. In the cell, this kind of misalignment of interests can occur when a sender gene is selfish, and would prefer to replicate itself at the expense of the rest of the genome. Such genes are termed "selfish elements," and come in a variety of forms, marked by decoupled replication from the rest of the genome (Burt and Trivers, 2006). In a signaling game, when there is such a conflict of interest, then the sender is expected to adopt a degree of deceptive signaling (Crawford and Sobel, 1982). Consistent with this, there are a range of selfish elements that utilize molecular deception, which implies that there is a cost to the host genome (Massey and Mishra, 2018). In addition, cancer

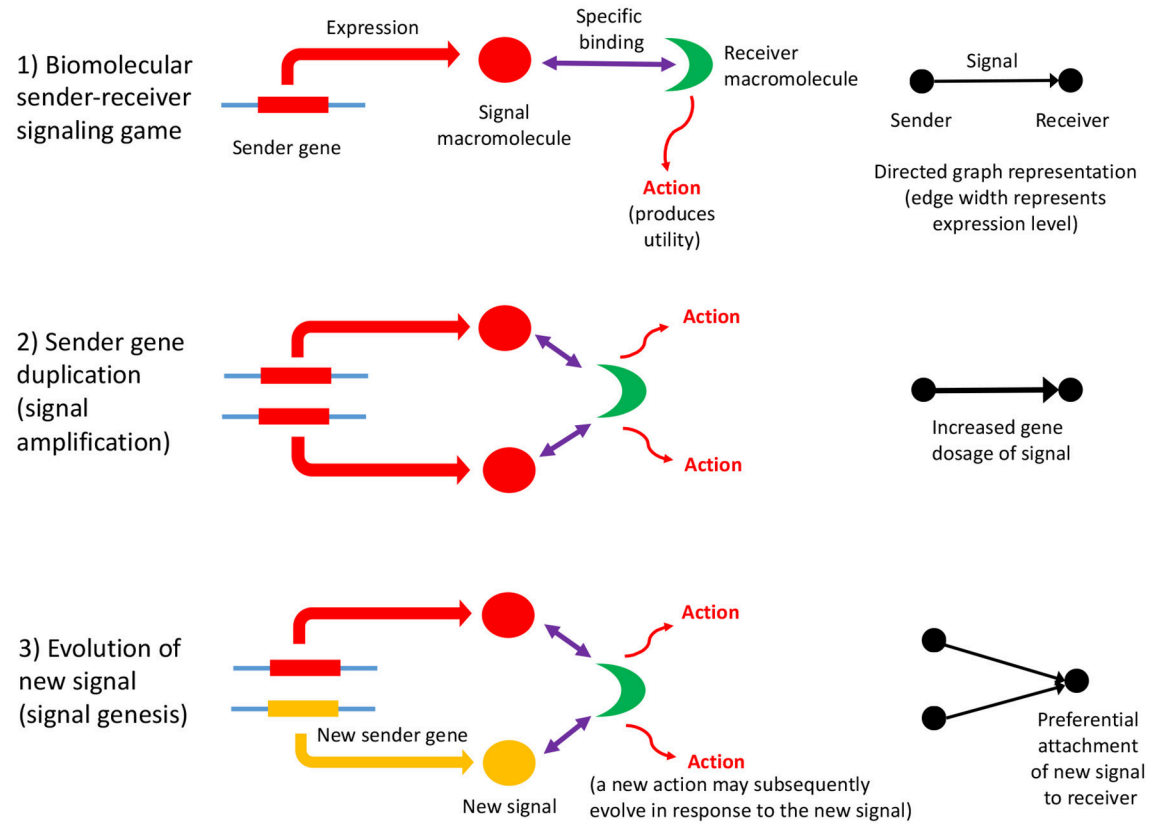


FIGURE 1 | The influence of information asymmetry on growth of a PPI network. Interactions between macromolecules are envisaged as a biomolecular signaling game whereby a sender gene expresses a macromolecule, the signal, that then binds specifically to a receiver macromolecule, which then undergoes an action (such as an enzymatic reaction, or conformational change), which produces utility (fitness). The signal consists of the three-dimensional conformation and physicochemical properties of the macromolecule (1). The sender gene may undergo duplication, which has a dosage effect on the expressed macromolecule, resulting in signal amplification (2). This mechanism is expected to lower the Shapley value of the gene players in the genome, as the signal is partially redundant and so inefficient. Subsequently, the sender gene duplicate may acquire a new function (evolve a new signal) although the majority would be expected to undergo pseudogenization (3). Both these scenarios represent the re-establishment of a Nash equilibrium. If a new signal macromolecule evolves, it is likely to bind to the same receiver macromolecule initially. This preferential attachment arises because gene duplicates have a tendency to bind to their original interaction partner initially, and then subsequently undergo interaction turnover (Zhang et al., 2005), and is illustrated to the right of the figure. A key problem is how a new action by the receiver arises as the result of the evolution of a new signal; the new action may co-evolve with the new signal, or may be necessary first before a new signal can evolve. The latter would imply that receiver gene duplication and action genesis facilitates the evolution of new signals and sender genes (an exception would be when there is a conflict of interest; here the sender is more likely to make the first move in evolving a novel deceptive signal, and then the receiver would respond with a better discriminative recognition mechanism). This key, and novel aspect of gene duplication might be deciphered via consideration of the topology of directed graph representations of biomolecular interactions as sender-receiver signaling games. Refinements to the illustrated scheme include situations where the original signal protein binds to a variety of receiver proteins, or where the gene that codes for the receiver protein undergoes duplication (**Figure 2**).

and pathogens also make widespread use of deceptive strategies at the molecular level, which is expected given their clearly opposed interests with the host (Massey and Mishra, 2018).

The importance of information asymmetry at the molecular level is manifold. Given that information asymmetry leads to the possibility of molecular deception, this means that in a biomolecular network, in principle honest and deceptive signals could be mapped as honest or deceptive biomolecular interactions, respectively. This viewpoint may have importance in better understanding of processes such as cancer, in which molecular deception plays a central role in its progression (section 4.3) where we also formulate open problem 4.H), as well as of the dynamics of persistent infections. Given the harmful effects of molecular deception, it is necessary to reduce information asymmetry, in order to promote cooperation between gene players, in the normal functioning of the genome. For instance, in the theory of incomplete contracts, a topic linked with the economics of information, reduction in information asymmetry reduces the likelihood of deception between parties, which consequently promotes trust (Devos et al., 2012) and so cooperation (Lorenz, 1999). Given this framework, one may suggest a form of "molecular trust" that is promoted when the information asymmetry between two gene players is reduced (effectively increasing transparency), with the effect of promoting utility (fitness) for both players, since deception is less likely to occur. One means to achieve this effect is by the use of costly signals, which are costly to produce and so are more likely to be honest (Veblen, 1899; Spence, 1973; Zahavi, 1975);

such signaling establishes "molecular trust" because mimicking the signals is expensive. In biomolecular terms, a costly signal is represented by the unique 3D conformation and physico-chemical properties of a macromolecule, which are difficult to imitate given its complexity.

When biomolecules are expressed from sender genes of an unknown type, identity signals are necessary, and so information asymmetry provides additional explanatory power for understanding the dynamics of molecular recognition. Biomolecules may be considered as belonging to two groups, namely, self and non-self, corresponding to cooperative members of the genome, or not, respectively. Self or non-self biomolecules might be equivalated to an in-group and out-group respectively, in sociological terms, and this view then might then imply some loose parallels between the dynamics of bio-molecular and social networks. In this context, it is of interest to consider how non-self gene players may become integrated into the cell and its biomolecular networks. This process may result from the endosymbiosis of a microbial genome, or the acquisition of plasmids. As the non-self genes evolve increasing cooperativity with the host genome over time, the occurrence of molecular deception is expected to reduce. This is because the level of deception is correlated with the level of conflict of interest (Martinez and Godfrey-Smith, 2016) with the host genome; the greater the misalignment of interests, the greater the level of deception that is expected from the non-self genes (Massey and Mishra, 2018).

Under the scenarios supported by a game theoretic framework, one may speculate how biomolecular networks may have originated. The very first biomolecular interactions in early life would have been characterized by molecular specificity, a distinguishing feature of biochemistry (Konnyu and Czaran, 2011). Molecular specificity arises when organic molecules reach a certain size, additional size being necessary to bind a smaller ligand. Molecular specificity is a form of recognition, which effectively allows verification of a ligand. Considering the ligand as a signal, then the macromolecule is the receiver, and the gene that produces the ligand can be considered the sender agent. The very first biomolecular network would have consisted of two nodes, sender and receiver, with the edge connecting the two representing the signal. As more biomolecular interactions evolved, the network increased in numbers of nodes and edges. Increases in organismal complexity may be viewed as an increase in the numbers of genes in the genome, but the numbers of biomolecular interactions has more explanatory power. Thus fully understanding the nature of these interactions and how they evolve is necessary for better understanding the emergent phenotype of an organism. In the genome of the ancestral life form, once a number of genes with separate function had evolved, it then would have become beneficial to evolve gene regulation. Therefore, genes with the dedicated function of regulating other genes in the genome would have arisen (transcription factors). The combination of regulatory and functional genes would have comprised the first GRN. Increases in organismal complexity have been facilitated by an increase in the complexity of the GRN (Burton, 2014).

Gene duplication, accompanied by the establishment of new biomolecular interactions, therefore is a fundamental evolutionary driver of organismal complexity (Lespinet et al., 2002), from the first life forms onward. Although the precise mechanism(s) of gene duplication remains to be established (Reams and Roth, 2015), some generalities may be made in terms of signaling games. The first step in the process of duplication of a sender gene may be viewed as one of signal enhancement. Because gene duplication results in gene dosage effects, it also results in amplification of the signal, the expressed gene product (resulting in weighted graphs—discussed in section 2). This strategy can be viewed as lowering the overall utility of the genome, given that there is a cost involved in producing excessive signal. It is, thus, expected to lower the Shapley value (Shapley, 1969) of the gene players that cooperate within the genome. This conflict is usually resolved when the duplicated gene becomes pseudogenized, the usual fate of gene duplicates (Innan and Kondrashov, 2010).

Subsequent to duplication, the gene duplicates will sometimes diverge in function, although the exact mechanism remains to be elucidated (Innan and Kondrashov, 2010). This process represents signal divergence if the gene is a sender gene, and action divergence if the gene codes for a receiver macromolecule. The genesis of a new sender gene with a new signal may then promote evolution of a novel action by the receiver macromolecule, potentially facilitating duplication of the receiver gene itself. Likewise, the duplication of a receiver gene may facilitate the diversification of macromolecular signals that interact with the two duplicated receiver macromolecules. The process modifies the GRN or PPI network in a non-obvious manner and it deviates considerably from the way evolution of random graphs is usually treated, following Erdös and Rényi, discussed in more detail in section 3 (Erdös and Rényi, 1959). These entail more complex random network evolutionary models (several of which are discussed in further detail in section 3).

Signal and action genesis via gene duplication may have features in common with a Pólya's urn model of signal genesis (Alexander et al., 2012) (Pólya's urn models are statistical models that involve sampling with replacement influenced by the identity of the sampled element. These models can lead to a "*rich get richer*" effect, of which "preferential attachment" is an example, discussed in more detail in subsection 3.2). In this model, reinforcement of signals (similar to reinforcement learning) may promote the invention of new synonyms. These considerations may provide parallels for how signals originate elsewhere, not dissimilar to how new words in a language can arise from existing words by a process of derivation (Cotterell et al., 2017). Mechanistic commonalities in the process of signal genesis in these diverse systems as exhibited in GRNs remain to be explored. These models hint at a possibly new, but universal form of "preferential attachment" that drives the variations in biomolecular networks as well as the selectivity in Darwinian evolution.

1.3. Network Topology, Evolution by Duplication, and Preferential Attachments

Consequently, the topology of gene networks is non-deterministic and yet not memoryless, since it must encode layers of ripples produced earlier via the dynamics of gene duplication (paralogs and orthologs), as amplified during the

network's history. Just as physicists infer the theories of origin of the universe from the cosmic background radiation, we expect to enrich our understanding of the origin of machinery of life (e.g., codon evolution, evolution of multicellularity, evolution of sex etc.) from a rigorous analysis of the signaling games and their equilibria, which has rippled through the extant biomolecular networks. Taking this analogy further, we observe that the ripples in gravitational waves have been proposed to reflect the existence of parallel universes, whose presence created asymmetries in the initial conditions, giving rise to filamentary structures in the visible universe (Hawking and Hertog, 2018) This comparison is inspired by the notion of a "protein big bang" from a single (or handful of) ur-protein(s) in the first complex life forms, evolving by gene duplication into the extant "protein universe," hinting at the information asymmetries fossilized in the GRN and PPI networks (Dokholyan et al., 2002).

Likewise, we point out that information asymmetry in macromolecular sender-receiver interactions may point to evolutionary paths that might have been abandoned unexplored; which may suggest new engineering approaches needed by synthetic biology, or in drug discovery, or immuno-therapy. Note that during the process of evolution of signaling, gene duplication and deletion contribute to a certain degree of non-determinism and "conventionality" to the Nash equilibria that stabilize and manifest as non-trivial anisotropies in gene network topology.

In summary, the process of gene duplication, tempered by signal and action genesis can be thought of as a driver of preferential attachment in shaping the topology of gene networks, in which information asymmetry between senders and receivers is expected to play an indelible role. **Figure 1** illustrates a basic mechanism whereby signal genesis may lead to preferential attachment during the growth of a PPI network. Topological features expected to hint at this process include: (i) the degree distribution, (ii) hierarchicity, (iii) assortativity and many others; they require powerful statistical and algebraic tools—covered in the later sections, where it is assumed that genome evolution is a complex process involving diverse groups of mutations such as insertions, deletions, conversions, duplications, transpositions, translocations, and recombinations, and that it is further affected by selective constraints and effective population size and other factors such as the environment. With recent understanding of large scale cellular networks (regulatory, metabolic, protein-protein interactions) one must now aim at investigation between the evolutionary rates of a gene mutations and its effects on the network topology using mathematical models and analytics (see Wagner, 1994). For instance, combining sequence analysis in a single genome and its close relatives, one can infer the rate and tempo of the evolutionary dynamics acting on the genome, and the resulting effects on the network's algebraic structures. We provide an example of how evolution by duplication leads to a preferential attachment mode of gene network growth in **Figure 2**, using the duplication of the p53 gene, and its paralogs p63 and p73—all transcription factors regulating pathways involved in related phenotypes of somatic or developmental surveillance and interacting with similar family of genes (e.g., MDM2 or MDMX), as illustration[1].

Note that these abstract models generate refutable hypotheses that need experimental verification and support from mechanistic explanations. However, unfortunately, the biochemical processes involved in the hypothesized preferential attachment dynamics are not fully understood. For example, the duplication processes are often driven by Non-Homologous End Joining (NHEJ), a pathway that repairs double-strand breaks in DNA. To guide repair, NHEJ typically uses short homologous DNA sequences called microhomologies, which are often present in single-stranded overhangs on the ends of double-strand breaks (Chang et al., 2017). When the overhangs are perfectly compatible, NHEJ usually repairs the break accurately. However, imprecise repair can lead to inappropriate NHEJ resulting in translocations, duplications, and rearrangements (Rodgers and McVey, 2016), which add to variational processes that are random but not memoryless. Perhaps some of such hypotheses may need to be carefully examined using cancer genome data such as The Cancer Genome Atlas (TCGA), and models of tumor progression. This analysis may also explain efficacy of certain therapeutic interventions in cancer as well as their failures via drug and immuno resistance.

2. NETWORK ANALYSIS

In this section, in order to address the potential impact of information asymmetry on network evolution, it is first necessary to discuss fundamentals of graphs (in particular directed and weighted graphs), a mathematical formalism used in the study of biomolecular networks, as well as other related important topics. Consider a set of entities, denoted V and a set of binary relations between the entities $E \subseteq V \times V$. When V denotes biomolecules and E denotes interactions between them (e.g., regulations, proximity, synteny, etc.), the resulting graph represents a biomolecular network. One important advantage of graphs is that they have an intuitive graphical representation. Such networks evolve over time with additions and deletions to the sets V and E. In order to create a bridge to algebraic approaches, we extend the standard combinatorial definition by endowing it with additional maps.

Formally, a graph is a pair of sets $G = (V, E)$ where V are the vertices (nodes, points) and $E \subseteq V \times V$ are the edges (arcs), respectively. When E is a set of unordered pair of vertices the graph is said to be undirected or simple. In a directed graph (which could result from information asymmetry, for example) $G = (V, E, o, t)$, E consists of an ordered set of vertex pairs, i.e., for each edge $e \in E$, $e \rightarrow (o(e), t(e))$ where $o(e)$ is called the origin of the edge e and $t(e)$ is called the terminus of the edge e (Serre, 1980; Biggs, 1993). A graph is weighted if there is a map (weighting function, $w : E \rightarrow R_+$) assigning to each edge a positive real-valued weight. Weighting can represent the strength of a signal in a sender-receiver interaction, for example.

[1] A mutation in MDM affects all p53, p63, and p73 allowing utility tradeoffs between fecundity (through decreased embryonic lethality) and cancer risks (through reduced somatic surveillance) in a population.

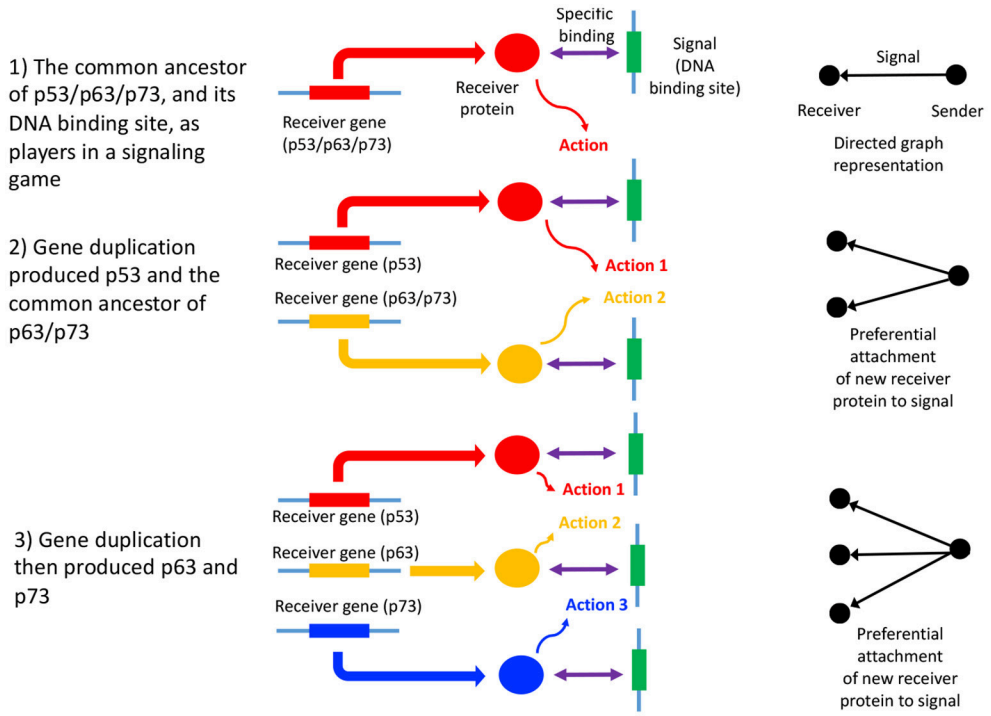


FIGURE 2 | Gene duplication of p53, p63, and p73 as a signaling game, and GRN growth. An illustrative example of a signaling games view of network growth is provided by the paralogs p53, p63, and p73, which code for transcription factors, p53 being of critical importance in many cancers (Joerger and Fersht, 2006). Here, p53 and the common ancestor of p63/p73 duplicated (2), followed by the duplication and divergence of p63 and p73 (Lu et al., 2009; Belyi et al., 2010) (3). The signal is the DNA binding site, while the receivers are the p53, p63, and p73 proteins (here the sender is the protein coding gene downstream of the DNA binding site). The receiver protein undergoes an action upon binding to the DNA binding site (the signal), which consists of the recruitment of additional transcription factors, and contribution to the assembly of the transcription initiation complex (Nogales et al., 2017). The gene products of p53, p63, and p73 mostly bind to the same DNA binding sites (Smeenk et al., 2008), thus each signal (and ultimately sender gene) has acquired two new binding partners, in addition to the original interaction with the gene product of the common ancestor of p53/p63/p73. This is a form of preferential attachment, which should influence network topology as the number of genes increase by duplication, as illustrated to the right of the figure. The signaling games perspective allows us to better understand scenarios where there is a conflict of interest between the genome, and a selfish entity such as a selfish element, a cancer or a virus. When there is a conflict of interest, a deceptive signal is expected to be emitted by the sender (Crawford and Sobel, 1982) (the selfish entity). Here, the DNA binding site of the selfish entity will mimic that of canonical DNA binding sites associated with normal cellular function, "tricking" a transcription factor to bind to it, and altering the transcription of the sender gene (or alternatively abolishing transcription factor binding). Examples include *cis*-regulatory mutations in cancer (Poulos et al., 2015).

If $G = (V, E, \cdot, \cdot)$ and $G' = (V', E', \cdot, \cdot)$ are two graphs such that $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$, then $G' \subseteq G$, G' is a *subgraph* of G. If $E' = E \cap (V' \times V')$ (E' contains every edge in $e \in E$ with $o(e), t(e) \in V'$) then G' is an *induced subgraph* of G. G' and G are *isomorphic* ($G' \equiv G$) if there is a bijection $f : V' \to V$ with $(u, v) \in E' \iff (f(u), f(v)) \in E, \forall u, v \in V'$.

2.1. Topological Properties

A network's properties are governed by its topology, such as the degree distribution, clustering coefficients, motifs, assortativity, etc. Comprehensive treatments for general networks can be found in Thulasiraman et al. (2015) and Loscalzo and Barabási (2016), and for more in-depth treatment regarding biomedical networks in Loscalzo et al. (2017). Here we discuss these properties in the context of biomolecuar networks, more specifically with respect to information asymmetry. The **Supplementary Material** contains a more complex combinatorial and algebraic graph theoretic approach.

Degree Distribution

The degree of a vertex v, $\deg(v)$, is the number of edges that connect the vertex with other vertices. In other words, the degree is the number of immediate neighbors of a vertex. In directed graphs the in-degree and out-degree of a vertex can be defined as the number of incoming and outgoing edges, respectively. Let n_k be the number of vertices of degree k and $|V| = N$, the total number of vertices in the graph and $|E| = M$, the total number of edges in the graph. Note that $\sum_k n_k = N$ and $\sum k n_k = \sum_{v \in V} \deg(v) = 2|E| = 2M$. The degree distribution is the fraction of vertices of degree k, $P(k) = n_k/N$, and two isomorphic networks will have the same degree distributions (though not necessarily the converse). Thus, the degree distributions can tell a great deal about the structure of a family of networks. For example, if the degree distribution is singly peaked, following the Poisson (or its Gaussian approximation) distributions, the statistical properties of the nodes can be described by the average degree $\langle k \rangle = \sum_k kP(k) = 2M/N$. The graph is said to be

sparse, if $\langle k \rangle = o(\log N)$ (or $M = o(N \log N)$). Biomolecular networks are usually sparse, which can be fruitfully exploited in their algorithmic analysis. We can talk of *typical* nodes of the networks as being those that have degree distribution as those within 1 to 2 standard deviations from the average, while, with probability decreasing exponentially, it is possible to find nodes with a degree much different from the average. While power-law degree distributions follow a completely different pattern: they are *fat-tailed*; the majority of the nodes have only a few neighbors, while many nodes have a relatively large number of neighbors. The highly connected nodes are known as *hubs*.

Distance Metrics

One of the most fundamental metrics is the *distance* on a graph. First we define a *walk* of length m in a graph G from a vertex u to v as a finite alternating sequence of vertices and edges $\langle v_0, e_1, v_1, e_2, \ldots, e_m, v_m \rangle$, such that $o(e_i) = v_{i-1}$ and $t(e_i) = v_i$, for $0 < i \leq m$, such that $u = v_0$ and $v = v_m$. Then the number of edges traversed in the shortest walk joining u to v is called the *distance* in G between u and v denoted by $d(u, v)$. If there is a walk from u to itself, then we say that the set of vertices (respectively edges) form a cycle. The smallest number of m edges in a walk from u to itself is called a cycle of length m. The girth $g(G)$, is the shortest cycle in G. A walk whose vertices are distinct is called a (simple) *path*.

The concept of a walk allows us to define other properties of the graph. A graph $G = (V, E, o, e)$ is said to be *connected*, if any two vertices are the extremities of at least one walk. The maximally connected subgraphs are called the *connected components* of G. A giant component is a connected component containing a significant fraction of the nodes. The maximum value of the distance function in a connected graph is called the *diameter* of the graph. Frequently real life networks have a small diameter and are said to exhibit the *small world phenomenon*. For many biomolecular networks the average distance between two nodes depends logarithmically on the number of vertices in the graph.

Additionally, a *complete graph* G is the undirected graph, in which each vertex is a neighbor of all other vertices; $\deg(v) = N - 1, \forall v \in V$; or equivalently, each distinct pair of vertices are connected (or are adjacent) by a unique edge. G is then denoted as K_N. A *clique* in an undirected graph is a subset of vertices such that its induced subgraph is complete. Additional combinatorial invariants of graphs useful in the analysis of networks can be defined (see **Supplementary Material** for details).

Expanding Constants

Let $G = (V, E, \cdot, \cdot)$ be an undirected graph. Then for all $F \subset V$, the *boundary* ∂F is the set of edges connecting F to $V \setminus F$. The *expanding constant*, or *isoperimetric constant* of X is defined as,

$$h(X) = \min_{\emptyset \neq F \subset V} \frac{|\partial F|}{\min\{|F|, |V \setminus F|\}}.$$

For a biomolecular network, then, the invariant $h(X)$ measures the quality of the network with respect to the flow of information within it, (e.g., via chemical reactions, or signaling). A larger $h(X)$ implies better expansion, faster mixing, faster partitioning, and many other related properties that may give the network a selective advantage.

Using various combinatorial algorithms devised for the study and analysis of biomolecular networks, one may compute $h(X)$ to determine their complexity. However, a precise characterization of $h(X)$ itself is an intractable (i.e., NP-complete) problem. Isoperimetric inequalities give bounds on $h(X)$ in terms of a related algebraic invariant, $\gamma(X)$ – called its *spectral gap*, determination of which has complexity $O(|V|)^c$, where c is at most 3; furthermore, $c = 1$ for many sparse graphs. We give isoperimetric bounds and results applicable to biomolecular networks in the **Supplementary Material**, where we also introduce a local Cheeger constant. We also introduce algebraic invariants in section 2.2.

Clustering and Clustering Coefficients

Biomolecular networks are modular, forming communities and hierarchies, likely to have been sculpted by EBD (Evolution by Duplication). To study these local structures in network science, one may perform *community* analysis, which aims to identify a group of nodes that have a higher probability of connecting to each other than to nodes from other communities [see for example (Pellegrini, 2019)]. These can be explained by our game theoretic formalism, and local Nash equilibria (see Massey and Mishra, 2018). Various notions such as k-cliques, k-clubs, and k-clans have been developed to detect communities, but they are ultimately closely connected to the problem of finding cliques and consequently, do not generally lend themselves to any reasonable algorithm other than brute-force enumeration. However, even detecting communities approximately may prove valuable for general evolutionary studies, since these

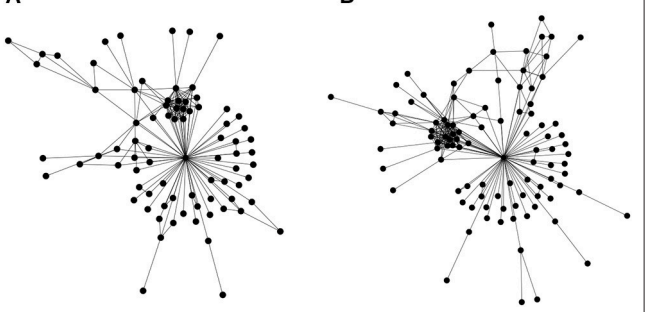


FIGURE 3 | Topological alignment of networks. Similar biomolecular networks could be topologically aligned and compared in order to express an evolutionary distance, which may then augment the traditional approaches of phylogenetic study. In order to account for the evolution by gene duplications, genes (or gene families) are to be identified and connected to their roles in biochemical pathways. Such an approach would lead to a program to understand the critical role of information asymmetries in driving evolution. Network alignment, a core problem in this program, is computationally intractable. To sharpen our intuition, we illustrate the problem using the social networks of the Gospels of Matthew and Luke. These networks represent social interactions between characters in the gospels of Matthew **(A)** and Luke **(B)**. These were chosen as a basic test for topological alignment procedures, given that they share a similar number of nodes, and the highly connected node of Jesus. A straightforward test for the efficacy of a topological alignment algorithm therefore constitutes aligning both networks and verifying that the Jesus node from both networks is matched. Edge lists for the two social networks may be found in the **Supplementary Materials**.

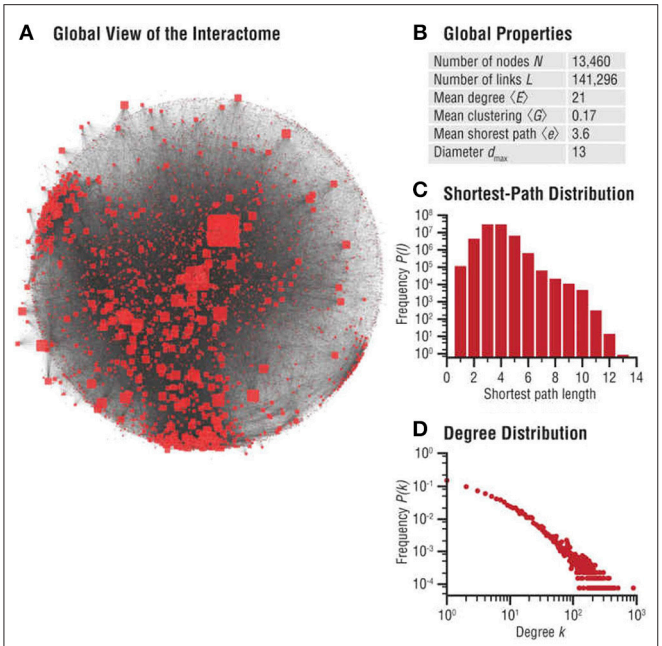


FIGURE 4 | Interactome networks used in the study of diseases. Undesirable interactions within a biomolecular network result in various disease states. Disease neighborhoods within the interactome can then be mapped to understand the progression of the disease [details can be found in Loscalzo et al. (2017)]. The progression of cancer has been studied using analysis of functionalization of oncogenes and dysfunctionalization of tumor suppressor genes via copy number fluctuations, but much more can be learned from the topological features of these genes in their interaction neighborhood. This illustration is from Figure 2.3 **(A–D)** of Menche and Barabási (2017). **(A)** Global map of the interactome, illustrating its heterogeneity. Node sizes are proportional to their degree, that is, the number of links each node has to other nodes. **(B)** Basic characteristics of the interactome. **(C)** Distribution of the shortest paths within the interactome. The average shortest path is $\langle d \rangle = 3.6$. **(D)** The degree distribution of the interactome is approximately scale-free (reproduced with permission from the publisher and authors of Menche and Barabási, 2017).

biomolecular network communities determine how specific biological functions are encoded in cellular networks—and are thus subjected to Darwinian selective pressure, since these players are likely to have formed communities in the first place to carry out specific cellular functions (see Hartwell et al., 1999), maximizing the utility of the cell. **Figure 4** highlights significant evidence that communities play an important role in human disease networks (see Loscalzo et al., 2017).

Usually a simpler approach is commonly employed and deals with the problem of *clustering* in a graph, which seeks to partition the graph into disjoint subgraphs such that nodes in each such subgraph are "closer" to the other nodes in the same subgraph, while they are "farther" from the nodes of other subgraphs. Hierarchical clustering algorithms have been developed to uncover communities (approximately) in polynomial time and depend upon the *similarity matrix* (x_{ij}), where the entry x_{ij} equals the distance between node i and node j. Among the classical algorithms are included those by Girvan and Newman (2002). Other related algorithms include those for random-walk betweenness and network centrality.

The *local clustering coefficient* captures the degree to which the neighbors of a given node link to each other. In general, for undirected graphs, the *local clustering coefficient* C_i of node i with degree k_i is defined as

$$C_i := \frac{L_i}{k_i(k_i-1)/2}$$

where the numerator L_i is the actual number of connections between k_i immediate neighbors of i, and the denominator is the number of connections if the neighbors formed a complete graph (i.e. a clique). Note that an undirected complete graph K_{k_i} of k_i nodes has $k_i(k_i-1)/2$ edges. Thus, a fully clustered node will have $C_i = 1$ and for completely isolated node $C_i = 0$. We can define the *(average) clustering coefficient* of the whole network with N nodes as

$$\langle C \rangle = \frac{1}{N}\sum C_i.$$

The clustering coefficients can be used to characterize a network's *modularity*, as discussed later (in section 3) in detail. For weighted graphs and directed graphs (as in the context of information asymmetry), a similar formalism is discussed in the **Supplementary Material**.

Subgraphs and Motifs

Biomolecular networks have been found to contain network *motifs*, representing elementary interaction patterns between small subgraphs that occur substantially more often than as predicted by a completely random network of similar size and connectivity. The presence of such motifs is usually explained by an evolutionary process that can quickly create (usually by a variation involving duplication) or eliminate (usually by a selection process that favors pseudogenization and complementation) regulatory interactions in a fast evolutionary time scale—relative to the rate at which individual genes mutate. It is usually hypothesized that the underlying evolutionary processes are convergent. Thus efficient algorithms to detect such motifs are important in the analysis of biomolecular networks. These algorithms focus on estimating how much more frequently a subgraph isomorphic to a motif graph (with n vertices and m edges) occurs relative to what would be expected by pure chance.

The number N_{mn} of subgraphs with n nodes and m interactions expected of a network of N nodes can be estimated from the two key topological parameters of a complex network—namely the power-law exponent β and the hierarchical exponent α as we discuss in Equations (1 and 2) below. In general the subgraph motifs can be classified in two types: Type I motifs are those where $(m - n + 1)\alpha - (n - \beta) < 0$, and type II subgraph motifs are those that satisfy the reverse inequality. One can determine their numbers N^I and N^{II} approximately as a function of $(m-n+1)\alpha-(n-\beta)$ and n_{max}, the degree of the most connected node in the network. One can show that $N^I >> N^{II}$. One can also show that the relative number of Type II subgraphs is vanishingly small compared to Type I.

2.2. Algebraic Invariants and Spectrum

The intuitive pictorial/combinatorial representation of graphs is an extremely useful aid to their understanding. However,

computing the topological properties of graphs combinatorially is computationally challenging especially when the size of the graph becomes large. As noted earlier, indeed, most combinatorial algorithms on biomolecular networks such as on PPI networks and GRNs are computationally complex problems (most of them fall in the NP-complete complexity class) (Karp, 2011). Therefore, in order to carry out any quantitative and computational analysis, graphs are better represented as algebraic objects. This representation allows us to use linear algebra and mathematical analysis techniques. The key to this representation is the adjacency matrix $A(G)$. It is defined as $\{0,1\}^{n\times n}$ matrix in which, $A_{ij} = 1$ if the vertices i and j are connected $[\exists e \in E, o(e) = i, t(e) = j]$ and 0 otherwise. The matrix is symmetric if the graph is undirected. For weighted graphs we can assign weights w_{ij} for existing edges. Networks that incorporate information asymmetry are directed, and the analysis becomes more complex. We refer to the **Supplementary Material** for this treatment.

Algebraic properties provide us with tools to deduce various properties of the biomolecular networks. In particular, the spectral representation of the graph is of importance for a number of applications such as graph classification, diffusion, expansion and mixing (see the **Supplementary Material**). We can think of the adjacency matrix A as operating on the space $V = C^n$ of complex n-tuples written as column vectors x,y as follows $Ax \rightarrow y$. It can be shown that there are directions left invariant in this space. That is to say, $Ax_i = \lambda_i x_i$ where λ_i are the eigenvalues and corresponding x_i the eigenvectors (spanning invariant directions) of the adjacency matrix for $1 \leq i \leq n$. The spectrum of the graph G is defined as the collection of eigenvalues of the adjacency matrix $\mathrm{Spec(G)} = \mathrm{Spec(A)} = \lambda_1, .., \lambda_n$. Naturally, if A is a real symmetric matrix, then the eigenvalues of A are real.

In particular, one algebraic invariant of the graph is the *spectral gap* $\gamma(G)$. It can be shown that the spectral gap gives excellent bounds on a combinatorial invariant, the Cheeger constant $h(G)$. Since information asymmetry leads to directed, weighted graphs, some of which are bipartite networks, we discuss these deeper algebraic analytics in the **Supplementary Material**.

3. NETWORK EVOLUTION

Starting with the seminal work of Erdös and Rényi (1959), a number of mathematical frameworks have been developed to model the "evolution" of graphs, covering the family of biomolecular networks. These frameworks may prove useful in explaining why most biological networks have certain non-obvious properties: namely, (i) The small world property; (ii) High clustering coefficients (varying with degree distribution); (iii) Emergence of "hubs." Such network models are ultimately expected to capture various observed properties of biomolecular networks, and the evolutionary trajectories leading up to them. The novel factor of information asymmetry, modeling genes as players, may also be incorporated, using the basic principles outlined in the Introduction, and **Figures 1**, **2**.

3.1. Random Network Models

Erdös and Rényi Model

The Erdös and Rényi model of random graphs [ER-graphs, denoted $G(n,p)$] is characterized by two parameters, the number of vertices in the network N and the fixed probability of choosing edges p (Erdös and Rényi, 1959). The graph G is generated by choosing N vertices and connecting each pair of vertices with probability p. The model yields a network with approximately $p\binom{N}{2} = O(pN^2)$ randomly distributed edges. The probability of choosing a specified graph G with N vertices and e edges is therefore $\binom{M}{e}p^e(1-p)^{M-e}$, where $M = \binom{N}{2}$ = the maximum number of possible edges connecting N vertices.

It can be shown that in such random graphs the average vertex degree is $\langle k\rangle = p(N-1) = O(pN)$. The diameter of such a graph is $d = \ln N/\ln\langle k\rangle \approx \ln N/(\ln N - ln(1/p))$ which is small compared to the graph size. Thus, random graphs exhibit "the small world property." The degree distribution for ER graphs is a binomial distribution $P[\deg(u) = k] = \binom{(N-1)}{k}p^k(1-p)^{N-k-1}$, which for large N (relative to $1/p$: where $N = \lambda/p$) converges to the Poisson distribution $P[\deg(u) = k] = e^{-\lambda}\frac{\lambda^k}{k!}$. Then the local clustering coefficient is $C_i = p$ is independent of the degree of the node and the average clustering coefficient $C = p/N$ scales with the network size. Therefore, the standard ER random model seems not to capture either the properties of degree distribution or the clustering coefficient of biomolecular networks.

Typically, an ER random graph model is used as a "null model" for the evolutionary process. However, while deviations from randomness are frequently used as evidence for the direct action of natural selection, often non-randomness may reflect neutrally generated (non-adaptive) emergent phenomena (Massey, 2015). We emphasize here that many topological features of biomolecular networks are unlikely to be directly selected for, but instead are a side-product of network growth, and decay, captured by the dynamics of edge and node addition and removal.

Small World Model

Biomolecular networks have features that are not captured by the Erdös and Rényi random graph model. As we have seen, random graphs have a low clustering coefficient and they do not account for the formation of hubs. To rectify some of these shortcomings, the *small world model* or popularly known as the *six degree of separation model* was introduced as the next level of complexity for a probabilistic model with features that are closer to real world networks (Watts and Strogatz, 1998; Watts, 1999). The evolution and dynamics of such networks have been discussed in detail (Watts, 2003), in particular in the diseases propagation literature (Dodds and Watts, 2005).

In this model, the graph G of N nodes is constructed as a ring lattice, in which, (i) first, *wire*: that is, connect every node to $K/2$ neighbors on each side and (ii) second, *rewire*: that is, for every edge connecting a particular node, with probability p reconnect it to a randomly selected node.

The average number of such edges is $pNK/2$. The first step of the algorithm produces local clustering, while the second dramatically reduces the distance in the network. Unlike

random graphs, the clustering coefficient of this network $C = 3(K-2)/4(K-1)$ is independent of the system size. Thus, the small world network model displays the small world property and the clustering of real networks, however, it does not capture the emergence of hubby nodes (e.g., p53 in biomolecular networks)(part of one of the eight open problems that we formulate in section 4).

3.2. Scale-Free Network Models

Most biomolecular networks are hypothesized to have a degree distribution, described as *scale-free*. In a scale free network the number of nodes n_k of degree k is proportional to a power of the degree, namely, the degree distribution of the nodes follows a *power-law*

$$n_k = k^{-\beta}, \tag{1}$$

where $\beta > 1$ is a coefficient characteristic of the network (Barabási and Albert, 1999). Unlike in random networks, where the degree of all nodes is centered around a single value – with the probability of finding nodes with much larger (or smaller) degree decaying exponentially, in scale-free networks there are nodes of large degree with relatively higher probability (*fat tail*). In other words, since the power low distribution decreases much more slowly than exponential, for large k (heavy or fat tails), scale-free networks support nodes with extremely high number of connections called "hubs." Power law distribution has been observed in many large networks, such as the Internet, phone-call maps, collaboration networks, etc. (Képès, 2007; Barabási, 2009; Loscalzo and Barabási, 2016). A caveat to these reports is that inappropriate statistical techniques have often been used to infer power law distributions, and alternative heavy tailed distributions may fit the data better (Clauset et al., 2009). However, the power law is a useful approximation that allows mechanisms of network growth to be explored, such as preferential attachment, discussed next, while the examination of alternative heavy tailed distributions is set as an Open Problem.

Preferential Attachment

The original model of *preferential attachment* was proposed by Barabási and Albert (1999). The scheme consists of a local *growth rule* that leads to a global consequence, namely a power law distribution. The network grows through the addition of new nodes linking to nodes already present in the system. There is higher probability to preferentially link to a node with a large number of connections. Thus, this rule gives more preferences to those vertices that have larger degrees. For this reason it is often referred to as the "rich-get-richer" or "Matthew" effect. This can be formulated as a game theoretic problem originating from information asymmetry and associated Nash equilibrium, discussed in the Open Problems.

With an initial graph G_0 and a fixed probability parameter p, the preferential attachment random graph model $G(p, G_0)$ can be described as follows: at each step the graph G_t is formed by modifying the earlier graph G_{t-1} in two steps – with probability p take a *vertex-step*; otherwise, take an *edge-step*:

(i) *Vertex step:* Add a new vertex v and an edge $\{u, v\}$ from v to u by randomly and independently choosing u proportional its degree;
(ii) *Edge step:* Add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their degrees.

That is, at each step, we add a vertex with probability p, while for sure, we add an additional edge. If we denote by n_t and e_t the number of vertices and edges respectively at step t, then $e_t = t+1$ and $n_t = 1 + \sum_{i=1}^{t} z_i$, where z_i's are Bernoulli random variables with probability of success $= p$. Hence the expected value of nodes is $\langle n_t \rangle = 1 + pt$.

It can be shown that exponentially (as t asymptotically approaches infinity) this process leads to a scale-free network. The degree distribution of $G(p)$ satisfies a power law with the parameter for exponent being $\beta = 2 + \frac{p}{2-p}$. Scale-free networks also exhibit *hierarchicity*. The local clustering coefficient is proportional to a power of the node degree

$$C(k) \approx k^{-\alpha} \tag{2}$$

where α is called the *hierarchy coefficient*.

This distribution implies that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs. In other words, it means that the level of clustering is much larger than that in random networks.

Consequently, many of the network properties in a scale-free network are determined by local structures—namely, by a relatively small number of highly connected nodes (hubs). A consequence of this structure of the scale-free network is its extreme robustness to failure, a property also displayed by biomolecular networks and their modular structures. Such networks are highly tolerant of random failures (perturbations); however, they remain extremely sensitive to targeted attacks.

Assortativity Network Model

Assortative mixing refers to the property exhibited by a preference of nodes to attach to similar (respectively, dissimilar) nodes; for example, high-degree vertices exhibit preference to attach to high-degree (resp. low-degree) vertices. Network models, discussed earlier and including the preferential attachment model, do not capture such important properties exhibited by real biomolecular networks (Girvan and Newman, 2002). Assortativity can be measured by the Pearson correlation coefficient r of degrees of linked nodes (Girvan and Newman, 2002). A positive correlation means connections between nodes of similar degree (assortativity) and a negative correlation means connections between nodes with different degree (disassortativity). Unlike technological networks and social networks (that show assortative mixing), biological networks appear to evolve in a disassortative manner.

GRNs are represented by directed graphs, and all biomolecular networks may be represented as directed graphs when the factor of information asymmetry is introduced (**Figures 1**, **2**). Assortative mixing can be generalized to directed biological graphs (Piraveenan et al., 2012). For directed networks two new measures, in-assortativity and the out-assortativity , can be defined measuring the correlation between the in-degree r_{in} and out-degree r_{out} of the nodes respectively. Biological

networks, which have been previously classified as disassortative, have been shown to be assortative with respect to these new measures. Also it has been shown that in directed biological networks, out-degree mixing patterns contain the highest amount of Shannon information, suggesting that nodes with high local out-assortativity (regulators) dominate the connectivity of the network (Piraveenan et al., 2012). The occurrence of assortativity in social networks has been attributed to a process of homophily [that is people tend to associate with others on the basis of ethnicity, religion, sports preferences etc. (McPherson et al., 2001; Newman, 2003a)]. The mechanisms that give rise to assortativity in biomolecular networks likely arises by a similar proximate mechanism of like nodes forming edges with like nodes, but the ultimate cause(s) remain unclear.

Duplication Model

Our earlier discussions suggest that biomolecular networks exhibit power-law degree distribution. However, unlike other complex networks, such as the Internet, the growth exponent of biomolecular networks typically falls into a lower range $1 < \beta < 2$, as opposed to $\beta \geq 2$. This difference has been suggested to have resulted from evolution by gene duplication dominating the evolutionary mechanism (Chung et al., 2003). We have already discussed the duplication phenomenon based on information asymmetry in GRNs in section 1. Various biomolecular networks have been studied using a partial duplication process, which proceeds in the following manner: Let the initial graph G_0 have N_0 vertices. In each step, G_t is constructed from its previous graph G_{t-1} as follows: A random vertex u is selected. Then a new vertex v is added in such a way that for each neighbor w of u, a new edge (u, w) is added with probability p. The process is then applied repeatedly. The full duplication model is simply the partial model with $p = 1$.

It has been shown that as the number N of vertices becomes infinitely large (as is the case for most biomolecular networks), the partial duplication model with selection probability p generates power-law graphs with the exponent satisfying the transcendental equation (Chung et al., 2003)

$$p(\beta - 1) = 1 - p^{\beta - 1},$$

whose solution determines the scale-free exponent β as a function of p. In particular, if $1/2 < p < 1$ then $\beta < 2$.

For illustrative purposes, we describe below an abstract gene network growth model incorporating the processes of gene duplication and deletion, as described above (Mishra and Zhou, 2004; Zhou, 2005). Using a Markov chain model the following features were investigated: (i) the origination of the segmental duplication; (ii) the effect of the duplication on the genome structure; and (iii) the role of duplication and deletion process in the genomic evolutionary distance. Unlike standard models of stationary Markov chain models, most processes in evolutionary biology belong to the group of non-stationary Markov processes, in which the transition matrix changes over time, or depends upon the current state.

This model results in the neutral emergence of scale-free degree distributions. It shows that the genomes of different organisms exhibit different network properties, likely reflecting differences in the rates of gene duplication and deletion (Mishra and Zhou, 2004). The additional factor of information asymmetry is likely to affect the nature of gene duplication in terms of gene identity and rate of duplication, and may provide additional explanatory power for differences in network properties. This analysis provides an example of how network topology can be used to provide insight into fundamental molecular evolutionary (neutral/Markov) processes in different species. Note that the model is relatively idealized, as it does not account for higher order interactions in a population involving: effective population size and allelic fixations; sex, diploidy, and sex-chromosomes (e.g., X and Y in mammals or W and Z in birds, etc.); surveillance and repair in somatic cells; embryonic lethality; homologous recombination, etc. The mathematical model explored here is kept simple to motivate the machinery from graph theory developed later.

Hierarchical Network Models

Another interesting model, introduced by Ravasz and Barabasi and dubbed the *hierarchical network model*, simulates the characteristics of many real life complex models and may be relevant. The resulting networks have modularity, a high degree of clustering, and the scale-free property. Modularity refers to the network phenomenon where many sparsely inter-connected dense subgraphs can be identified—"one can easily identify groups of nodes that are highly interconnected with each other, but have only a few or no links to nodes outside of the group to which they belong" (from Ravasz and Barabási, 2003).

A generative process for a hierarchical network model may be described as follows: For instance, consider an initial network H_0 of c fully interconnected nodes (e.g., $c = 5$). As a next step, create $(c - 1)$ replicas of this cluster H_0 and connect the peripheral nodes of each replica to the central node of the original cluster to create H_1 with c^2 (e.g., $c^2 = 25$) nodes. This step can be repeated recursively and indefinitely, thereby for any k steps the number of nodes generating the graph H_k with c^{k+1} nodes. If the central nodes of H_0 is called a *hub* and other nodes *peripheral*, then each recursion replicates additional copies of hubs and peripheral nodes.

One can carry out a recursive analysis and show that one obtains a power-law (i.e., scale-free) network with exponent $\beta = 1 + \frac{\ln(c)}{\ln(c-1)}$. The local clustering coefficients (for the hub-nodes) follow $C(k) \approx \frac{2}{k}$. Also, one can show that this duplication feature of the evolutionary process leads to hierarchical behavior of the network. The resulting networks are expected to be fundamentally modular, in other words, the network can be seamlessly partitioned into collections of modules where each module performs an identifiable task, separate from the function(s) of other modules. One can also show that the average clustering coefficient on N nodes at any given stage is about $C = 0.7419282..$ (for $c = 4$), $C = 0.741840$ (for $c = 5$), and a constant for a fixed c, independent of N (see Ravasz and Barabási, 2003, and for exact computations Noh, 2003).

4. OPEN PROBLEMS AND FUTURE CHALLENGES

The study of biomolecular networks is still a relatively young field and has thus far focused on a mechanistic perspective. As we begin to explore biomolecular networks from a more involved evolutionary view point, we encounter a large array of promising areas of investigation—most of which focus on how information asymmetries among the gene players ultimately sculpt the information flow, as necessary for an organism to navigate in a complex and fluctuating environment. Molecular evolution has classically been concerned with the dualism of selection and neutrality, however here we have highlighted a third important component, information asymmetry, and suggest a series of Open Problems that may help to begin to better understand its impact. The traditional approaches of phylogenetic study may be applied here, but examining specifically the family of species-specific biomolecular networks. Thus, mathematically we would need the networks to be aligned, motifs to be mapped to each other and network-distances to be correlated to deep evolutionary time. In order to account for the evolution by duplications, *orthologs* and *paralogs* of a gene (or gene families) are to be identified and connected to their roles in biochemical pathways. Ultimately, this analysis could be targeted at extracting the origin of various information-asymmetric signaling games and how they are stabilized in their Nash equilibria.

Key questions include whether signaling game characteristics differ between species. For example, species may differ in their average sender/receiver ratio, and the average complexity of signals produced (which may be indicated by protein size, variability in expression, and degree of post-translational regulation). Such differences may be linked to organismal complexity, variability in the environment and multicellularity. In so doing an overarching picture of how information is gathered from the environment, and how it is shared and distributed amongst gene players might be intimated. In particular, at its core this program requires an explanation of how features of genome evolution and structure might be algorithmically inferred from a network science perspective, as follows.

4.1. Algorithmic Complexity Issues

A key problem central to this program would be in detecting isomorphism mappings among pairs of graphs or subgraphs, a problem of infeasible algorithmic complexity (assuming $P \neq NP$). We start with a discussion of these issues and cite heuristics that can tame the problem, albeit computing the solutions approximately.

Intractability: NP-Completeness

Many combinatorial optimization problems seem impossible to solve except by brute-force searches evaluating all possible configurations in the search space. They belong to a complexity class called NP-complete and include such problems as whether a graph has a clique of size k. Since finding certain recurrent motifs in a class of networks shares many computational characteristics of the clique problem and since it could be central to discovering important evolutionary signatures (e.g., EBD), it seems unlikely that it would be possible to characterize the evolutionary trajectories precisely—especially when the number of genes involved are in the thousands. See the **Supplementary Material** for additional discussions on graph representations and to derive their algebraic invariants, that provide bounds on complexity of algorithms possibly leading to excellent approximate results in the study of sparse complex networks (see Chung, 1997; Chung and Lu, 2006).

Problem 4.A *Classify various computational problems involved in detecting evolutionary trajectories of biomolecular networks and characterize their algorithmic complexity.*

Problem 4.B *Explore PTAS (Polynomial Time Approximation Schemes) for these problems—Especially when the graphs satisfy certain sparsity, modularity and/or hierarchy properties.*

Algebraic Approximation

As described earlier, many interesting topological features of a graph can be computed efficiently (on both sequential and parallel computers) from their descriptions in terms of adjacency matrices. The resulting spectral methods have found recent applications in complex networks (e.g., communication, social, Internet) (see Spielman, 1996, 2018; Chung, 1997, 2010; MacKay, 2003; Spielman and Teng, 2004, 2011, 2013, 2014; Chung and Lu, 2006). These methods are efficient (linear time complexity) for sparse graphs, whose number of edges is roughly of the same order as the number of vertices. Thus, they are well suited to biomolecular networks (for example for clustering, community detection, hubs, robustness, assortative mixing, spreading and mixing, closeness, isomorphism, among others).

Thus, spectral graph theory may be expected to have many applications in the analysis of biomolecular networks, most prominently, in clustering, graph similarity, and graph approximation, but also in smoothing analysis and sparsification. One can envisage that many, if not most, classical network algorithms in biomolecular networks can be made faster by spectral methods. Indeed, since most biomolecular networks are sparse—both in terms of sparse connections, and in precise algebraic sense (see the **Supplementary Material**), these algorithms likely lead to linear time algorithms. The smoothing analysis methods, as well as sparsification approximations are worth exploring in these contexts.

Another fruitful direction is in parallelizing these algorithms. As an illustration, in several studies of biomolecular networks it would be useful to identify when two networks X_1 and X_2 are "close." We may wish to say that two networks are close if $Spec(X_1)$ and $Spec(X_2)$ are close—a computational problem that is polynomially computable (and efficiently parallelizable) (see Spielman and Teng, 2013). We can now give a mathematical formulation of this closeness, which can also be incorporated into phylogenetic studies. These biomolecular networks may be annotated with weights that are linear or quadratic approximation of relations, as common in these studies. These analyses may identify sub-networks that have been influenced by EBD, in concert with selection.

Problem 4.C *Classify various algebraic problems involved in detecting evolutionary trajectories of biomolecular networks and characterize their ability to approximate. Explore their practical implementations on sequential and parallel computers.*

4.2. Design Principles via Motif Analysis

The study of Systems Biology postulates that there are important design principles of biological circuits that provide a great deal of insight. The connections of gene and protein interaction networks are assumed to provide the necessary robustness and control to achieve cellular function in the face of chemical noise. However, it remains unclear how random variations alone provide such robustness. A possible explanation may come from a game-theoretic model that leads to stable equilibria and is expected to have precipitated from duplication of genes, interactions, and motifs. In addition, in principle, the dynamics of biomolecular sender-receiver signaling games should be reflected in network topologies, and so give rise to particular motifs. While the specific types of motifs expected to be observed remains to be developed further, some general principles can be identified. As discussed in section 1, the dynamics of signal genesis are driven by gene duplication, which affects overall network topology, in terms of the degree distribution. However, subgraphs consist of groups of senders and receivers, which likely have a related role in the cell, this may be tested by approaches outlined by Dotan-Cohen et al. (2009). The topology of these subgraphs contain localized motifs, which again reflect the addition and deletion of sender and receiver genes. The impact of information asymmetry is expected to lie in the Nash equilibria and associated utilities of sender-receiver interactions, which should be an influence on whether a new biomolecular interaction is established, or not.

Machine Learning

The biomolecular networks of interest are derived from highly noisy data e.g., CHIP-Chip, CHIP-Seq (for GRN), or co-localization or two-hybrid (for PPI) and consequently, the inferred edges of the network may miss certain genuine interactions or include several spurious interactions. Various machine learning algorithms (with false discovery rates, control, and regularization techniques) have been devised in order to improve the accuracy of such models. Biomolecular networks from related species (with ortholog and paralog analysis) are often combined to improve the accuracies and cross-validate results. The accuracies may be further ascertained via various local properties.

One important local property of networks is determined by so-called network motifs, which are defined as recurrent and statistically significant sub-graphs or patterns. Thus, network motifs are sub-graphs that repeat themselves in a specific network or even among various networks. Each of these sub-graphs, defined by a particular pattern of interactions between vertices, may reflect a framework in which particular functions are achieved efficiently. Indeed, motifs are of notable importance largely because they may reflect functional properties. They have recently gathered much attention as a useful concept to uncover structural design principles of complex networks. Although network motifs may provide a deep insight into the network's functional abilities, their detection is computationally challenging. Thus an important challenge for both experimental and computational scientists would be to study the evolutionary dynamics starting with the experimental data *ab initio*, as well as in improving the accuracy and efficiency of both the experimental and algorithmic techniques simultaneously.

Problem 4.D *Classify the species distributions of the different forms of heavy tailed distributions (e.g., power law, exponential, power law with exponential decay, lognormal), in different types of biomolecular network, and infer the mechanistic causes during network growth, and ultimate molecular evolutionary origins.*

Problem 4.E *Characterize the motifs in the biomolecular networks of closely related species starting with the noisy experimental data. Explain the structure of the motifs via their effect on the information flow. For instance, one may focus on DOR (Dense Overlapping Regulons) motifs and how they might have evolved from a simpler ancestral regulon (Alon, 2006).*

Problem 4.F *Study Subgraph Isomorphism Algorithms (and heuristics) for sparse graphs and identify special cases most suitable for studying evolutionary trajectories, while relating them to biomolecular design principles.*

Network Alignment

Critical to the evolutionary studies, described above, is the topic of network alignment and subsequent network tree building, which may be used for the comparative approach, between species-specific networks. Networks may be aligned in a pairwise fashion to calculate similarity, and from this a distance matrix is calculated, and used for the construction of a network tree, showing the relationships between multiple networks. For example, in the case of meta-metabolic networks, such studies will reveal relationships between the meta-metabolic networks of different microhabitats. A plausible prediction is that the network tree should show convergent evolution in microbial communities from microhabitats with similar conditions (e.g., anaerobic habitats). Thus this approach could lead to a tool to study convergent evolution of microbial community structure in similar habitats (Goldford et al., 2018). The signaling games perspective promises a more complete view of the cooperation, and conflict, that is present in all microbial communities, and is expected to be reflected in the structure of meta-metabolic networks. In particular, cooperation will be indicated by honest signals, whereas conflict by the occurrence of deceptive signals, which are expected to include molecular mimics.

From an algorithmic point of view, one may employ any of the three types of network alignment approaches:

1. where node identity is known;
2. where node similarity can be determined (based on sequence similarity for example); and
3. where node identity is unknown, here only network topology is used for alignment.

The first is a straightforward edge alignment. However, a refined expression is required that incorporates similarities

in edge widths in addition to the basic edge alignment (presence/absence of common edges between networks). Most effort in bioinformatics has gone into the second type network alignment, where there is partial information regarding node identity (for example Kalaev et al., 2005; Pinter et al., 2005). There do exist some first generation heuristics that utilize the third type of alignment approach (only topology) (Kuchaiev and Przulj, 2011), but the underlying graph isomorphism problem is known to be #P-complete. But these heuristics, as would be expected, do not work well—a straightforward test for this problem is applying them to align the social networks of the Gospels of Luke and Matthew (**Figure 3**)—the Jesus node should always align, as it is rather obvious topologically; but often leads to failure.

Problem 4.G *Classify and characterize the graph alignment algorithms.*

4.3. Somatic Evolution and Cancer

Network analysis is used in disease studies, but there have been more focused studies with applications to disease processes in cancer. In **Figure 4** we show part of an interactome network useful in deciphering aberrant interactions in diseases (Figure 2.3 from Loscalzo et al., 2017). Cancer is a complex disease, but governed by somatic genomic evolution, as propelled by mutation. Thus as a consequence, GRNs may be used to better understand cancer susceptibility, map its progression, design better tailored therapies, and better understand the evolution of endogenous anti-cancer strategies. Cancer genes are often network hubs (Karimzadeh et al., 2018), as they are often involved in critical developmental pathways. But a better network analysis will shed light on many natural questions: Why is it so? How does this come about from the process of network growth over evolutionary time? What clues do they provide to understand the somatic evolution in cancer and its progression?

During cancer progression, the disease reduces a cell's healthy genome into an aberrant mutant, where cancer eventually leads to metastasis, ultimately resulting in death of the patient. The healthy cells in the patient may be thought to possess a normal network, that is a gene network that engenders health and well-being. Cancer progression is reflected by a dynamic change of the normal network into an aberrant network. The aberrant network manifests itself by tumorigenesis, and finally metastasis. There is a substantial literature enumerating the identity of oncogenes and tumor suppressor genes, which aberrantly gain function (e.g., amplification of copy number) or lose function (e.g., deletion in copy number, hemi- or homo-zygously), respectively. They modify the cell biology of cancer progression, effected via the dynamics of GRN and PPI networks in cancer progression—all remain to be fully characterized.

Figure 2 shows a simple model for how the evolution of p53 and its paralogs may affect GRN topology; such molecular evolutionary information-asymmetric signaling games approaches may help to better understand the motifs associated with oncogenes in GRNs. An additional important factor in cancer is the pervasive occurrence of molecular deception (Bhatia and Kumar, 2013). From a signaling games perspective, the use of deception is consistent with cancer's conflict of interest with somatic cells. The identity of deceptive macromolecular signals may be incorporated into the network, potentially shedding a novel light on the mechanism of carcinogenesis. The genesis of deceptive signals therefore is expected to impact and drive carcinogenesis, with the level of deception increasing as the cancer progresses, and as its conflict with the soma intensifies. Of interest is the question whether there is an identifiable phase transition in network topology associated with metastasis. Taming this deception should therefore constitute a key counter-strategy in combating cancer, and is currently represented by the use of immunotherapy approaches (Zhang and Chen, 2018), although the game theoretical underpinning of these techniques has not been appreciated.

An additional factor to understanding this biology are copy number variants (CNVs)—types of gene mutations where a number of large sections of genomic DNA may be duplicated (or deleted), resulting in dosage effects of the resident gene sequences, which are exactly duplicated (or deleted). The numbers of CNVs can commonly vary substantially within a population, and have been shown to have significant roles in the propensity to develop cancer (Krepischi et al., 2012). An increase in the number of CNVs would have the effect of enhancing the weight of an edge, which represents the interaction of the CNV gene product with its macromolecular binding partner. Such a network variant represents an increased disposition to develop cancer, and can be understood as occupying a position in "network space" (the space of all possible network topologies) in greater proximity to an aberrant network, than a normal network.

Problem 4.H *Study Cancer progression models in terms of GRN's and identify the role of* driver *and* passenger *genes in the somatically evolving networks, and the number and distribution of deceptive signals.*

4.4. Gene Regulation and 3D Networks

The origin and development of GRNs from a signaling games perspective is discussed in the Introduction. However, GRNs typically do not take into account 3D spatial orientation, and this provides a more complete view of gene regulation. Recent work has outlined the importance of three-dimensional proximity of genes to genes on other chromosomes, in addition to their immediate neighborhood on their own chromosome (Li et al., 2018). This effect implies that gene proximity and spatial relationships within the nucleus can be meaningfully represented as a network. Such a network would be comprised of two types of edge: (1) linear distance on the same chromosome (centimorgans), (2) physical distance with genes on other chromosomes (nanometers). Such networks may be termed "3D gene orientation networks."

Gene regulation and co-regulation may be better understood by the construction and analysis of 3D gene orientation networks. This is because the proximity of regulatory modules to a gene has an influence on gene expression. Most genes have a regulatory region 5′ of the transcription start site, the promoter. In addition, regulatory enhancers and other regulatory elements

may be located distant from the gene, generally on the same chromosome (Gondor and Ohlsson, 2018). It is thought that the bending and juxtaposition of chromosomes within the nucleus may bring such elements into physical proximity to the gene (Gondor and Ohlsson, 2018). Clearly, the physical distance, and frequency with which the element is brought into contact with the gene will influence the nature of its regulatory input. Using 3D gene orientation networks, additional information may be incorporated into edges, such as whether physical proximity is static, or has movement. If there is movement, this may be coordinated (or not) with other regulatory elements affecting the same gene. Likewise, interactions with regulatory elements may show some coordination between genes. A signaling games aspect is incorporated by considering the regulatory elements as signals, the gene that is regulated as the sender, and DNA binding proteins that bind to the regulatory elements as receiver molecules, this scheme is illustrated in **Figure 2**.

Problem 4.I *Describe the Gene Duplication process and their signaling game utilities in terms of the genome's 3D structure.*

4.5. Generalization of Genetic Variations

This paper describes an idealized picture: it describes a canonical gene regulation network and variations affecting the associated (single) genome, among which gene duplication has taken a lion's share of the focus. This picture needs to be generalized to consider an ensemble of genomes, and variations to the implied ensemble of genetic networks, which can vary based on additional intra-genome variations: e.g., horizontal gene transfer, reverse transcription and recombination, but also due to effects such as cell-fusion and endosymbiosis and effect of population sizes (e.g., in allelic fixation, for instance in sex chromosomes). Mathematically, the implied models of family of graphs would be significantly complex and may require theories from large networks and graph limits to understand the asymptotic properties. We leave these and associated algorithmic questions as topics of future research.

Problem 4.J *Adding genome duplication and fusion, gene transfer, gene conversion, endosymbiosis, sexual recombination, fixation etc. to describe evolution of an ensemble of GRNs.*

5. CONCLUSION

Here, we have outlined graph theoretical approaches that may reveal some novel aspects of the molecular evolutionary process, incorporating the understudied factor of information asymmetry, whose effect may become manifest at the level of the phenome. Further work is required to link the diverse features of network topology with network evolution and growth. While the evolutionary aspects shaping individual gene-gene interactions has been addressed by geneticists and molecular evolutionists, we believe that a synthesis entailing a multi-disciplinary effort combining game theory, graph theory, and algebraic/statistical analysis will provide a more informative omnigenic model of gene interactions, in contrast to the traditional homogenic view. Given our view that biomolecular networks may be modeled using evolutionary game theory, and given that evolutionary game theoretical approaches have been used in the study of social networks, we expect that some surprising similarities and convergences between the topologies of the two might be observed. Finally, we note that the field of statistics gained impetus from the consideration of biological problems, from workers such as Fisher, Haldane, Rao, Wright, Kimura, Crow, and others, and so we suggest that consideration of the open problems listed here might also lead to a similar development of new mathematics.

6. BIBLIOGRAPHIC NOTES

We recommend the following articles for further reading: (Albert and Barabási, 2002; Barabási et al., 2002, 2003, 2004; Farkas et al., 2002; Schwartz et al., 2002; Barabási, 2003; Chung and Lu, 2004, 2006; Candia et al., 2008; Goh and Barabási, 2008; Vazquez et al., 2008; Davis et al., 2010; Song et al., 2010; Liu et al., 2013; Janwa and Rangachari, 2015). For other important sources (especially with respect to directed graphs), we refer to Newman and Watts (1999), Newman (2001, 2003b,c,d, 2004, 2006, 2010), Girvan and Newman (2002), Meyers et al. (2006), Moore et al. (2006), Clauset et al. (2009), Karrer and Newman (2010), Newman et al. (2011), Zhang et al. (2016, 2017). For evolution of networks (see for example Sharan et al., 2005; Mazurie et al., 2010). For bipartite networks (Janwa and Lal, 2003; Hø holdt and Janwa, 2012). For Spectral methods (Cvetković et al., 1980; Lubotzky et al., 1988; Lubotzky, 1994, 2012; Chung, 1997; Davidoff et al., 2003; Sarnak, 2004; Chung and Lu, 2006; Spielman and Teng, 2011; Janwa and Rangachari, 2015).

AUTHOR CONTRIBUTIONS

BM conceived of and structured the presented ideas at a high level. SM and BM developed the biological theories. HJ, BM, and JV developed the computational, quantitative, and mathematical theories. All authors discussed the open problems and contributed to the final manuscript.

FUNDING

This work was supported by National Science Foundation Grants CCF-0836649 and CCF-0926166, and a National Cancer Institute Physical Sciences-Oncology Center Grant U54 CA193313-01 (to BM).

ACKNOWLEDGMENTS

We acknowledge our colleagues in UPR and NYU, who have generously provided many constructive criticisms.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00240/full#supplementary-material

REFERENCES

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* 74, 47–97. doi: 10.1103/RevModPhys.74.47

Alexander, J. M., Skyrms, B., and Zabell, S. L. (2012). Inventing new signals. *Dyn. Games Appl.* 2, 129–145. doi: 10.1007/s13235-011-0027-2

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca Raton, FL: Chapman and Hall; CRC.

Barabási, A.-L. (2003). "Emergence of scaling in complex networks," in *Handbook of Graphs and Networks, From Genome to the Internet* ed H. G. Schuster (Weinheim: Wiley-VCH), 69–84.

Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science* 325, 412–413. doi: 10.1126/science.1173299

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512. doi: 10.1126/science.286.5439.509

Barabási, A.-L., Dezső, Z., Ravasz, E., Yook, S.-H., and Oltvai, Z. (2003). "Scale-free and hierarchical structures in complex networks," in *Modeling of Complex Systems Vol. 661 of AIP Conference Proceedings* (Melville, NY: American Institute of Physics), 1–16.

Barabási, A.-L., Oltvai, Z. N., and Wuchty, S. (2004). "Characteristics of biological networks," in *Complex Networks*, Vol. 650, Lecture Notes in Physics, eds. E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Berlin: Springer), 443–457.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Phys. A* 311, 90–614. doi: 10.1016/S0378-4371(02)00736-7

Belyi, V. A., Ak, P., Markert, E., Wang, H., Hu, W., Puzio-Kuter, A., et al. (2010). The origins and evolution of the p53 family of genes. *Cold Spring Harb. Perspect. Biol.* 2:a001198. doi: 10.1101/cshperspect.a001198

Bhatia, A., and Kumar, Y. (2013). Cellular and molecular mechanisms in cancer immune escape: a comprehensive review. *Expert Rev. Clin. Immunol.* 10, 758–762. doi: 10.1586/1744666X.2014.865519

Biggs, N. (1993). *Algebraic Graph Theory*. Cambridge Mathematical Library. Cambridge: Cambridge University Press.

Burt, A., and Trivers, R. (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Cambridge, MA: Harvard University Press.

Burton, Z. F. (2014). The old and new testaments of gene regulation. *Transcription* 5:e28674. doi: 10.4161/trns.28674

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A* 41:224015. doi: 10.1088/1751-8113/41/22/224015

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., and Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell. Biol.* 18, 495–506. doi: 10.1038/nrm.2017.48

Chung, F. (2010). Graph theory in the information age. *Notices Am. Math. Soc.* 57, 726–732.

Chung, F., and Lu, L. (2004). "The small world phenomenon in hybrid power law graphs," in *Complex Networks*, Vol. 650, Lecture Notes in Physics, eds E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (Berlin: Springer), 89–104.

Chung, F., and Lu, L. (2006). *Complex Graphs and Networks*. Vol. 107, CBMS Regional Conference Series in Mathematics. Washington, DC; Providence, RI: American Mathematical Society.

Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks. *J. Comput. Biol.* 10, 677–687. doi: 10.1089/106652703322539024

Chung, F. R. K. (1997). *Spectral Graph Theory*. Vol. 92, CBMS Regional Conference Series in Mathematics. Washington, DC; Providence, RI: American Mathematical Society.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* 51, 661–703. doi: 10.1137/070710111

Cotterell, R., Vylomova, E., Khayrallah, H., Kirov, C., and Yarowsky, D. (2017). "Paradigm completion for derivational morphology," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagan), 714–720.

Crawford, V. P., and Sobel, J. (1982). Strategic information transmission. *Econometrica* 50, 1431–1451. doi: 10.2307/1913390

Cvetković, D. M., Doob, M., and Sachs, H. (1980). *Spectra of Graphs*. Vol. 87, Pure and Applied Mathematics. New York, NY: Academic Press Inc.; Harcourt Brace Jovanovich Publishers.

Davidoff, G., Sarnak, P., and Valette, A. (2003). *Elementary Number Theory, Group Theory, and Ramanujan Graphs*. Vol. 55, London Mathematical Society Student Texts. Cambridge: Cambridge University Press.

Davis, D. A., Chawla, N. V., Christakis, N. A., and Barabási, A.-L. (2010). Time to CARE: a collaborative engine for practical disease prediction. *Data Min. Knowl. Discov.* 20, 388–415. doi: 10.1007/s10618-009-0156-z

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., and Hahn, M. W. (2018). The evolution of mammalian gene families. *PLoS ONE* 1:e85. doi: 10.1371/journal.pone.0000085

Devos, J., Van Landeghem, H., and Deschoolmeester, D. (2012). "Chapter 11: The theory of the lemon markets in is research," in *Information Systems Theory: Explaining and Predicting our Digital Society, Vol. 1*, eds Y. K. Dwivedi, M. R. Wade, and S. L. Schneberger (New York, NY: Springer), 213–230.

Dodds, P. S., and Watts, D. J. (2005). A generalized model of social and biological contagion. *J. Theor. Biol.* 232, 587–604. doi: 10.1016/j.jtbi.2004.09.006

Dokholyan, N. V., Shakhnovich, B., and Shakhnovich, E. I. (2002). Expanding protein universe and its origin in from the biological big bang. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14132–14136. doi: 10.1073/pnas.202497999

Dotan-Cohen, D., Letovsky, S., Melkman, A. A., and Kasif, S. (2009). Biological process linkage networks. *PLoS ONE* 4:e5353. doi: 10.1371/journal.pone.0005313

Erdös, P., and Rényi, A. (1959). On random graphs. *Public. Math.* 6, 290–297.

Farkas, I., Derényi, I., Jeong, H., Néda, Z., Oltvai, Z. N., Ravasz, E., et al. (2002). Networks in life: scaling properties and eigenvalue spectra. *Phys. A* 314, 25–34. doi: 10.1016/S0378-4371(02)01181-0

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Goh, K.-I., and Barabási, A.-L. (2008). Burstiness and memory in complex systems. *Europhys. Lett. EPL* 81:48002. doi: 10.1209/0295-5075/81/48002

Goldford, J. E., Lu, N., Bajic, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., et al. (2018). Emergent simplicity in microbial community assembly. *Science* 361, 1390–1396. doi: 10.1126/science.aat1168

Gondor, A., and Ohlsson, R. (2018). Enhancer functions in three dimensions: beyond the flat world perspective. *F1000Research* 7:681. doi: 10.12688/f1000research.13842.1

Govindarajan, S., and Goldstein, R. A. (1997). Evolution of model proteins on a foldability landscape. *Proteins* 29, 461–466. doi: 10.1002/(SICI)1097-0134(199712)29:4<461::AID-PROT6>3.0.CO;2-B

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402(6761 Suppl.), C47–C52. doi: 10.1038/35011540

Hawking, S. W., and Hertog, T. (2018). A smooth exit from eternal inflation? *arXiv:1707.07702*. doi: 10.1007/JHEP04(2018)147

Hø holdt, T., and Janwa, H. (2012). Eigenvalues and expansion of bipartite graphs. *Des. Codes Cryptogr.* 65, 259–273. doi: 10.1007/s10623-011-9598-6

Huang, L., Liao, L., and Wu, C. H. (2017). Evolutionary analysis and interaction prediction for protein-protein interaction network in geometric space. *PLoS ONE* 12:e0183495. doi: 10.1371/journal.pone.0183495

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689

Janwa, H., and Lal, A. K. (2003). On Tanner codes: minimum distance and decoding. *Appl. Algeb. Engrg. Commun. Comput.* 13, 335–347. doi: 10.1007/s00200-003-0098-4

Janwa, H., and Rangachari, S. (2015). "Ramanujan graphs and their applications," in *Research Monograph* (San Juan).

Joerger, A. C., and Fersht, A. R. (2006). The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches. *Annu. Rev. Biochem.* 85, 375–404. doi: 10.1146/annurev-biochem-060815-014710

Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. (2005). Networkblast: comparative analysis of protein networks. *Bioinformatics* 24, 594–596. doi: 10.1093/bioinformatics/btm630

Karimzadeh, M., Jandaghi, P., Papadakis, A. I., Trainor, S., Gonzalez-Porta, M., Scelo, G., et al. (2018). Aberration hubs in protein interaction networks highlight actionable targets in cancer. *Oncotarget* 9, 25166–25180. doi: 10.18632/oncotarget.25382

Karp, R. M. (2011). Heuristic algorithms in computational molecular biology. *J. Comput. Syst. Sci.* 77, 122–128. doi: 10.1016/j.jcss.2010.06.009

Karrer, B., and Newman, M. E. J. (2010). Message passing approach for general epidemic models. *Phys. Rev. E* 82:016101. doi: 10.1103/PhysRevE.82.016101

Képès, F. E. (2007). *Complex Systems and Interdisciplinary Science.* Biological Networks. Singapore: World Scientific.

Konnyu, B., and Czaran, T. (2011). The evolution of enzyme specificity in the metabolic replicator model of prebiotic evolution. *PLoS ONE* 6:e20931. doi: 10.1371/journal.pone.0020931

Krepischi, A. C. V., Pearson, P. L., and Rosenberg, C. (2012). Germline copy number variations and cancer predisposition. *Fut. Oncol.* 8:681. doi: 10.2217/fon.12.34

Kuchaiev, O., and Przulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* 27, 1390–1396. doi: 10.1093/bioinformatics/btr127

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059. doi: 10.1101/gr.174302

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

Li, Y., Hu, M., and Shen, Y. (2018). Gene regulation in the 3D genome. *Hum. Mol. Genet.* 27, R228–R233. doi: 10.1093/hmg/ddy164

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2013). Observability of complex systems. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2460–2465. doi: 10.1073/pnas.1215508110

Lorenz, E. (1999). Job market signaling. *Cam. J. Econ.* 23, 301–315.

Loscalzo, J., and Barabási, A.-L. (2016). *Network Science, 1st Edn.*. Cambridge: Cambridge University Press.

Loscalzo, J., Barabási, A.-L., and Silverman, E. K. E. (eds.). (2017). *Network Medicine: Complex Systems in Human Disease and Therapeutic, 1st Edn.* Cambridge, MA: Harvard University Press.

Lu, W.-J., Amatruda, J. F., and Abrams, J. M. (2009). p53 ancestry: gazing through an evolutionary lens. *Nat. Rev. Cancer* 9, 758–762. doi: 10.1038/nrc2732

Lubotzky, A. (1994). *Discrete Groups, Expanding Graphs and Invariant Measures*. Vol. 125, Progress in Mathematics. Basel: Birkhäuser Verlag.

Lubotzky, A. (2012). Expander graphs in pure and applied mathematics. *Bull. Amer. Math. Soc.* 49, 113–162. doi: 10.1090/S0273-0979-2011-01359-3

Lubotzky, A., Phillips, R., and Sarnak, P. (1988). Ramanujan graphs. *Combinatorica* 8, 261–277. doi: 10.1007/BF02126799

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. New York, NY: Cambridge University Press.

Martinez, M., and Godfrey-Smith, P. (2016). Common interest and signaling games: a dynamic analysis. *Philos. Sci.* 83, 371–392. doi: 10.1086/685743

Massey, S. E. (2015). Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life* 5, 1301–1332. doi: 10.3390/life5021301

Massey, S. E., and Mishra, B. (2018). Origin of biomolecular games: deception and molecular evolution. *J. R. Soc. Interface* 15:20180329. doi: 10.1098/rsif.2018.0429

Mazurie, A., Bonchev, D., Schwikowski, B., and Buck, G. A. (2010). Evolution of metabolic network organization. *BMC Syst. Biol.* 4:59. doi: 10.1186/1752-0509-4-59

McCloskey, D., Palsson, B. O., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* 9:661. doi: 10.1038/msb.2013.18

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* 27, 415–444. doi: 10.1146/annurev.soc.27.1.415

Menche, J., and Barabási, A.-L. (2017). "Chapter 2. Introduction to network analysis," in *Network Medicine: Complex Systems in Human Disease and Therapeutic, 1st Edn*, eds J. Loscalzo, A.-L. Barabási, and E. K. Silverman (Cambridge, MA: Harvard University Press).

Meyers, L. A., Newman, M. E. J., and Pourbohloul, B. (2006). Predicting epidemics on directed contact networks. *J. Theor. Biol.* 240, 400–418. doi: 10.1016/j.jtbi.2005.10.004

Mishra, B., and Zhou, Y. (2004). *Models of Genome Evolution.* Natural Computing Series, Lecture Notes in Computer Science. (Berlin: Springer Verlag), 287–304.

Moore, C., Ghoshal, G., and Newman, M. E. J. (2006). Exact solutions for models of evolving networks with addition and deletion of nodes. *Phys. Rev. E* 74:036121. doi: 10.1103/PhysRevE.74.036121

Newman, M., Barabasi, A.-L., and Watts, D. J. (2011). *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* 98, 404–409. doi: 10.1073/pnas.98.2.404

Newman, M. E. J. (2003a). Assortative mixing in networks. *Phys. Rev. Lett.* 89, 758–762. doi: 10.1103/PhysRevLett.89.208701

Newman, M. E. J. (2003b). Mixing patterns in networks. *Phys. Rev. E* 67:026126. doi: 10.1103/PhysRevE.67.026126

Newman, M. E. J. (2003c). "Random graphs as models of networks," in *Handbook of Graphs and Networks, From the Genome to the Internet* (Weinheim: Wiley-VCH), 35–68.

Newman, M. E. J. (2003d). The structure and function of complex networks. *SIAM Rev.* 45, 167–256. doi: 10.1137/S003614450342480

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74:036104. doi: 10.1103/PhysRevE.74.036104

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.

Newman, M. E. J., and Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Phys. Lett. A* 263, 341–346. doi: 10.1093/acprof:oso/9780199206650.001.0001

Newman, M. E. J. (2004). "Who is the best connected scientist? A study of scientific coauthorship networks," in *Complex Networks*, Vol. 650 of Lecture Notes in Physics, eds E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (Berlin: Springer), 337–370.

Nogales, E., Louder, R. K., and He, Y. (2017). Structural insights into the eukaryotic transcription initiation machinery. *Ann. Rev. Biophys.* 46, 59–83. doi: 10.1016/S0375-9601(99)00757-4

Noh, J. D. (2003). Exact scaling properties of a hierarchical network model. *Phys. Rev. E* 67:045103. doi: 10.1146/annurev-biophys-070816-033751

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

Pellegrini, M. (2019). "Community detection in biological networks," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schonbach (Oxford: Academic Press), 978–987.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. (2005). Alignment of metabolic networks. *Bioinformatics* 21, 3401–3408. doi: 10.1093/bioinformatics/bti554

Piraveenan, M., Prokopenko, M., and Zomaya, A. Y. (2012). On congruity of nodes and assortative information content in complex networks. *Netw. Heterog. Media* 7, 441–461. doi: 10.3934/nhm.2012.7.441

Poulos, R. C., Sloane, M. A., Hesson, L. B., and Wong, J. W. H. (2015). The search for *cis*-regulatory driver mutations in cancer genomes. *Oncotarget* 6, 32509–32525. doi: 10.18632/oncotarget.5085

Ravasz, E., and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Phys. Rev. E* 67:026112. doi: 10.1103/PhysRevE.67.026112

Reams, A. B., and Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *CSH Perspect. Biol.* 7:a016592. doi: 10.1101/cshperspect.a016592

Rodgers, K., and McVey, M. (2016). Error-prone repair of dna double-strand breaks. *J. Cell. Physiol.* 231, 15–24. doi: 10.1002/jcp.25053

Sarnak, P. (2004). What is. . . an expander? *Notices Amer. Math. Soc.* 51, 762–763.

Schuster, P., Fontana, W., Stadler, P., and Hofacker, I. (1994). From sequences to shapes and back: a case-study in RNA secondary structures. *Proc. R. Soc. Lond. B* 255, 279–284. doi: 10.1098/rspb.1994.0040

Schwartz, N., Cohen, R., ben Avraham, D., Barabási, A.-L., and Havlin, S. (2002). Percolation in directed scale-free networks. *Phys. Rev. E* 66:015104. doi: 10.1103/PhysRevE.66.015104

Serre, J.-P. (1980). *Trees*. Berlin: Springer-Verlag. Translated from the French by John Stillwell.

Shapley, L. S. (1969). "A value for n person games," in *The Shapley Value,* ed A. Roth (Cambridge: Cambridge University Press), 31–40.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., et al. (2005). Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1974–1979. doi: 10.1073/pnas.0409522102

Smeenk, L., van Heeringen, S. J., Koeppel, M., van Driel, M. A., Bartels, S. J. J., Akkers, R. C., et al. (2008). Characterization of genome-wide p53-binding sites upon stress binding. *Nucleic Acids Res.* 36, 3639–3654. doi: 10.1093/nar/gkn232

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science* 327, 1018–1021. doi: 10.1126/science.1177170

Spence, M. (1973). Job market signaling. *Q. J. Econ.* 87, 355–374. doi: 10.2307/1882010

Spielman, D. (2018). *Spectral Graph Theory and Its Applications*. Available online at: http://www.cs.yale.edu/homes/spielman

Spielman, D. A. (1996). Linear-time encodable and decodable error-correcting codes. *IEEE Trans. Inform. Theory* 42(6 Pt 1), 1723–1731. doi: 10.1109/18.556668

Spielman, D. A., and Teng, S.-H. (2004). Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *J. ACM* 51, 385–463. doi: 10.1145/990308.990310

Spielman, D. A., and Teng, S.-H. (2013). A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM J. Comput.* 42, 1–26. doi: 10.1137/080744888

Spielman, D. A., and Teng, S.-H. (2014). Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.* 35, 835–885. doi: 10.1137/090771430

Spielman, D. A., and Teng, S.-H. (2011). Spectral sparsification of graphs. *SIAM J. Comput.* 40, 981–1025. doi: 10.1137/08074489X

Taylor, P. D., and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Math. Biosci.* 40, 145–156. doi: 10.1016/0025-5564(78)90077-9

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu. Rev. Cell Dev. Biol.* 31, 399–428. doi: 10.1146/annurev-cellbio-100913-012908

Thulasiraman, K., Arumugam, S., Brandstadt, A., and Nishizeki, T. (2015). *Handbook of Graph Theory, Combinatorial Optimization, and Algorithms*. CRC Computer and Information Science Series. Boca Raton, FL: Taylor & Francis; Chapman & Hall.

Vazquez, A., de Menezes, M. A., Barabási, A.-L., and Oltvai, Z. N. (2008). Impact of limited solvent capacity on metabolic rate, enzyme activities, and metabolite concentrations of *S. cerevisiae* glycolysis. *PLoS Comput. Biol.* 4:e1000195. doi: 10.1371/journal.pcbi.1000195

Veblen, T. (1899). *The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions*. New York, NY: Macmillan.

Wagner, A. (1994). Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. U.S.A.* 91, 4387–4391. doi: 10.1073/pnas.91.10.4387

Watts, D. J. (1999). *Small Worlds: The dynamics of networks between order and randomness*. Princeton Studies in Complexity. Princeton, NJ: Princeton University Press.

Watts, D. J. (2003). *Six Degrees*. New York, NY: W. W. Norton & Co. Inc.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). ipath2.0: interactive pathway explorer. *Nucleic Acids Res.* 39, W412–W415. doi: 10.1093/nar/gkr313

Zahavi, A. (1975). Mate selection - a selection for a handicap. *J. Theor. Biol.* 53, 205–214. doi: 10.1016/0022-5193(75)90111-3

Zhang, H., and Chen, J. (2018). Current status and future directions of cancer immunotherapy. *J. Cancer* 9, 1773–1781. doi: 10.7150/jca.24577

Zhang, P., Moore, C., and Newman, M. E. J. (2016). Community detection in networks with unequal groups. *Phys. Rev. E* 93:012303. doi: 10.1103/PhysRevE.93.012303

Zhang, X., Moore, C., and Newman, M. E. J. (2017). Random graph models for dynamic networks. *Eur. Phys. J. B* 90:200. doi: 10.1140/epjb/e2017-80122-8

Zhang, Z., Luo, Z. W., Kishino, H., and Kearsey, M. J. (2005). Divergence pattern of duplicate genes in protein-protein interactions follows the power law. *Mol. Biol. Evol.* 22, 501–505. doi: 10.1093/molbev/msi034

Zhou, Y. J. (2005). *Statistical Analyses and Markov Modeling of Duplication in Genome Evolution*. Ph.D. thesis, New York University (NYU).

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.

Copyright © 2019 Janwa, Massey, Velev and Mishra. 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) and the copyright owner(s) 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 Applied Mathematics and Statistics

REVIEW
published: 12 April 2019
doi: 10.3389/fams.2019.00018

Metabolic Games

Taneli Pusa [1,2,3]*, *Martin Wannagat* [1,2] *and Marie-France Sagot* [1,2]*

[1] *INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin, France,* [2] *Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, CNRS, Université de Lyon, Université Lyon 1, Villeurbanne, France,* [3] *Department of Computer, Automatic and Management Engineering, Sapienza University of Rome, Rome, Italy*

Metabolic networks have been used to successfully predict phenotypes based on optimization principles. However, a general framework that would extend to situations not governed by simple optimization, such as multispecies communities, is still lacking. Concepts from evolutionary game theory have been proposed to amend the situation. Alternative metabolic states can be seen as strategies in a "metabolic game," and phenotypes can be predicted based on the equilibria of this game. In this survey, we review the literature on applying game theory to the study of metabolism, present the general idea of a metabolic game, and discuss open questions and future challenges.

Keywords: metabolic modeling, flux balance analysis, evolutionary game theory, microbial interactions, metabolic networks

OPEN ACCESS

Edited by:
Bud Mishra,
New York University, United States

Reviewed by:
Priyanka Baloni,
Institute for Systems Biology (ISB), United States
Ranjan K. Dash,
Medical College of Wisconsin, United States

***Correspondence:**
Taneli Pusa
henri.pusa@inria.fr
Marie-France Sagot
marie-france.sagot@inria.fr

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *07 December 2018*
Accepted: *27 March 2019*
Published: *12 April 2019*

Citation:
Pusa T, Wannagat M and Sagot M-F (2019) Metabolic Games. Front. Appl. Math. Stat. 5:18. doi: 10.3389/fams.2019.00018

1. INTRODUCTION

Metabolic networks have become a standard model in computational biology and high quality genome-scale reconstructions are now available for a wide range of micro-organisms as well as of some eukaryotes. Often the ultimate aim of these models is phenotype prediction, which means predicting from the genome how an organism would behave in a given environment. In this context, constraint-based methods, most prominently Flux Balance Analysis (FBA), have a proven track record in accurately predicting the metabolic behavior of single organisms [1–5].

FBA relies on assumptions about the underlying optimization principles guiding metabolic behavior, and biomass yield relative to nutrient intake is often chosen as the target of maximization. While this assumption is often justified when considering single species systems, it becomes troublesome if one wishes to model several species at the same time [6]. Simple optimization is usually not enough, because competition and interspecies interactions complicate the situation considerably. Formulating a "common goal" for a community of organisms can only be done *ad-hoc* [7–9]. Moreover, there are situations where even in single species communities, selection can be unfavorable to optimal choices such as maximizing efficiency in nutrient use [10–12].

Game theory is a branch of applied mathematics originally developed to describe and reason about situations where two or more rational agents, the "*homo economicus*," are faced with choices and have potentially conflicting goals [13]. All participants want to maximize their own well-being, but are doing so taking into account that everyone else is doing the same. Thus paradoxical, suboptimal, outcomes are possible and even common. *Evolutionary* game theory was born out of the realization that rational choice can be replaced by natural selection: in the course of evolution the strategy (phenotype) that would "win" the game would prevail by simply proliferating more successfully thanks to its success in the "game" [14, 15].

It turns out that phenotype prediction in the context of metabolic networks is exactly the type of problem that evolutionary game theory was meant to answer: given a set of choices (as defined by a metabolic network reconstruction), what will be the actual metabolism observed? In other words, if we culture a set of organisms together in a given medium, which are the phenotype(s) that emerge as winners?

In this review, we seek to provide a short introduction to both evolutionary game theory and its use in the context of metabolic modeling. We first present the relevant preliminaries and introduce the idea of a metabolic game. We then further expand on the idea by reviewing work done on the topic so far. Finally, we discuss these ideas and contemplate on future prospects.

We wish to call attention to the fact that our focus here is strictly on the idea of using the principles of game theory to reason about metabolic networks. While some papers that address this topic have been included for the sake of completeness, we decided to omit part of the related literature to keep the scope of this review under control. For previous reviews discussing the use of game theory in the context of micro-organisms with slightly different emphases, see [11, 16–18].

Game Theory

The main concepts that compose a *game* are a set of players, a set of actions for each player, and a payoff function. The players are the participants in the interaction. In the simplest case, they can be interchangeable, meaning they all have the same set of available actions and the same payoff function. A set of actions defines the choice that each player faces and can correspond for example to the expressed phenotype. Finally, the payoff function determines the outcome for each player in each scenario, that is, a combination of actions chosen by the players.

The simplest game is the 2-player, 2-strategy matrix game. If the players are interchangeable, it can be expressed concisely by the payoff matrix:

	A	**B**
A	a	b
B	c	d

where A and B denote the actions and the entries are payoff values for the row player. For example, if the row player plays A and the column player B, the payoffs for the row and column players are b and c, respectively.

Some of these games have become famous and the actions and payoffs can be given generic interpretations, usually denoted by:

	C	**D**
C	R	S
D	T	P

where C stands for "cooperation" and D for "defection," and the payoffs, denoted by their initials, are known as "Temptation," "Reward," "Punishment," and "Sucker's payoff." If $T > R > P > S$, the game is a Prisoner's Dilemma (PD). It corresponds to a situation where the players would both be better off cooperating, but because they will always have the incentive to defect, they end up choosing this inferior outcome, hence the "dilemma."

A common way to analyse games is using a *solution concept*. A solution is a state of the game (in other words, a configuration of actions/strategies) that can be reasonably assumed to follow from choices made based on some underlying logic. Arguably the two most well-known examples—as well as the ones most often encountered in the context of evolutionary game theory—are the Nash equilibrium [19, 20] and the Evolutionarily Stable Strategy (ESS) [14]. In a Nash equilibrium, all strategies are chosen in such a way that no player has an incentive to unilaterally change theirs. An ESS is a strategy such that if adopted by every member of a population, a small minority of players using any other strategy cannot invade. In the Prisoner's Dilemma, for both players to choose D is a Nash equilibrium of the game: D dominates the other action in all scenarios ($T > R$ and $P > S$) [13]. In this case, it is also an ESS.

If the payoffs are switched so that $T > R > S > P$, the game is called Hawk-Dove or Snowdrift (SD). In contrast to the PD, in this situation it is still better to cooperate even if one's partner fails to do so. Here the Nash equilibrium is to choose an action opposite of one's opponent. If mixed strategies are allowed, meaning a player can choose its action probabilistically, we have a mixed Nash equilibrium where both players follow the same strategy of choosing C with some probability (or a portion of the time). This is also an ESS, and can be interpreted as a population of individuals that comprises a mix of C- and D-players.

The simple matrix game can be readily extended in two ways. First, the number of strategies can be increased, effectively increasing the dimensions of the matrix. This in general leads to no extra complications apart from eventually requiring computational tools for the analysis of the equilibrium structure (see [21–23]). Second, the number of payoff matrices and players can be increased. In other words, players are no longer interchangeable and there can be more than two parties in the interaction. In general, matrix games with more than two players are much harder to analyse than simpler games [24].

The most prominent multiplayer game is the Public Goods. It can be thought of as an extension of the PD to more than two players. In the simplest form, n players each choose whether or not to make a contribution to the common good. The contribution has a cost c for the individual, and yields a benefit $r \cdot c$ that is distributed evenly amongst the group. If $r/n < 1$, no one will contribute, even though everyone would be better off with all members making the contribution. However, this is true with the kind of linear benefits in the simplest game, but not necessarily in a more general case where the benefit acquired can be a non-linear function of the contributions. In Archetti and Scheuring [25] and Broom and Rychtář [24], it has been argued that in real situations, benefits are usually not linear, but for example saturating.

In principle, any population dynamics model can be used in combination with a game by simply making the growth rate a function of the payoff. Obviously this requires defining with whom and how the game is played. For example, it can be assumed that the population is well-mixed, so that players encounter different types of opponents in a random fashion according to their prevalence. The payoff that an individual is obtaining at a given moment is thus the expectation, calculated over the different possible encounters. The most commonly used formulation is the replicator equation [26, 27]. It describes the dynamics of the frequencies of strategies as:

$$\frac{dn_i}{dt} = n_i(E_i - \bar{E}), \tag{1}$$

where n_i is the relative density of strategy i in the population, E_i is its expected payoff, and $\bar{E}$ is the average (expected) payoff of the population.

Adaptive dynamics is a framework that combines the ecological and evolutionary time scales to study how strategies will evolve under natural selection [28–30]. Under the assumptions that changes in strategy undergo gradual mutation so that each new genotype changes the phenotype only slightly, and that the mutations occur rarely enough for the ecological and evolutionary dynamics to be separable, adaptive dynamics offers a view beyond simply reasoning about the stable points of evolution such as an ESS. For example, an ESS can be unattainable through gradual mutations. Adaptive dynamics can also explain how evolution toward higher fitness in a homogeneous population can lead to diversity, or so-called branching of the evolutionary tree [30].

Metabolic Games

A metabolic network (see **Figure 1**) is represented by the so-called stoichiometric matrix S with m rows corresponding to the number of metabolites and n columns to reactions. The standard steady state assumption:

$$S \cdot \mathbf{v} = 0 \tag{2}$$

expresses the condition that any flux vector $\mathbf{v}$ must render the net production of all internal metabolites zero. However, this still leaves the state of the metabolism largely undetermined. In fact, the steady state condition merely defines the set of possible metabolic strategies available to the organism, comprising all the different pathways at its disposal. The question then is what is the choice made by the organism: which pathways it chooses to activate.

"Choice" here obviously refers to that made by natural selection. Thus, the question could be put more appropriately as: given the environment, as determined by both abiotic factors as well as the surrounding members of the same and of different species, what is the best response to this environment. Again, "best response" refers to the ability to persist in competition with other members of the community, generally referred to as *fitness*.

In FBA, the metabolic state is inferred through straightforward optimization (for an illustration of the FBA workflow, see **Figure 2**). A standard choice is the flux through a biomass reaction. While this is often referred to as growth rate, strictly speaking it corresponds to growth yield [6, 10] (see also section 2). It can be seen as fitness maximization in isolation. As argued throughout this paper, this might not correspond to the strategy of choice if the surrounding community is taken into account. However, if the metabolic state is already sufficiently specified through additional constraints, this growth yield maximization can still be used to determine the fitness *given that specific choice*. We can thus define a metabolic game: the players are cells, actions are the different metabolic states available to them, and payoffs are calculated using FBA with additional constraints specifying the states chosen in each combination of actions. A schematic representation of this idea is given in **Figure 3**.

Consider the toy-model example in **Figure 4**. The simple network represents a situation where an organism has two options for a primary nutrient: A and B. We assume that in order to efficiently utilize whichever nutrient is chosen, the cell has to specialize. Thus, uptake of both A and B is not feasible. Furthermore, nutrient A is superior, yielding 3 units of the biomass precursor M_1 per one unit of A as opposed to a yield of only 2 for B. Thus, following standard FBA, we would conclude that by choosing to uptake A at the maximum rate 1, the cell maximizes its biomass production ($v_{Biom} = 3$).

We now add a social interaction component by assuming that the presence of A in the environment is limited (for the sake of simplicity we assume that B is abundant). This can be modeled by a simple 2-player matrix game with two actions: to only uptake A and to only uptake B (denoted by "MS1" and "MS2" resp.). Should both players choose A as their nutrient, the maximum uptake rate is halved ($v_{TA} \leq 0.5$), reflecting a scarcity of the compound. In this case, maximizing biomass production only yields 1.5 (assuming that $v_{TB} = 0$).

The pure NE of the game is the "anti-coordination" scenario where the players choose differing strategies. The mixed NE and as well as the ESS is a strategy where both players' choice can be expressed as 2/3MS1 + 1/3MS2, meaning that nutrient A is chosen two-thirds of the time. This can be interpreted as a stochastic strategy where the cell switches from pure MS1 to pure MS2 randomly. The equilibrium of the replicator dynamics (Equation 1) corresponds to the mixed NE and the ESS but with a different interpretation: in a well-mixed population, two-thirds of the cells will use A while the remaining uses B.

The most complete realization to date of this formalism was presented in Zomorrodi and Segre [23]. Indeed, Zomorrodi and Segrè construct different metabolic strategies by setting selected fluxes to zero to simulate knock-outs, and forcing the excretion of "leaky" metabolites that can be taken up by neighboring cells. Payoffs are obtained by maximizing the biomass flux for both genotypes in each pairwise interaction. Together, these define a 2-player matrix game with 2 or more actions (genotypes). To determine which genotype(s) are able to persist, the authors search for Nash equilibria and ESSs using the replicator equation.

As a proof of concept, Zomorrodi and Segrè apply their framework to study invertase production in the yeast *Saccharomyces cerevisiae*: in order to grow on sucrose, the yeast needs to hydrolyse the sugar molecule. Because invertase is a surface enzyme, much of the resulting monosaccharides leak out. Because producing invertase is costly, it constitutes a public good. This cost is modeled by reducing the ATP production of invertase-producers. It was found that depending on how much of the sugar leaks out and on the cost of producing the enzyme, three different payoff schemes are possible: Prisoner's Dilemma, Snowdrift, and Mutually Beneficial.

The authors also studied amino acid mediated ecological interactions in *Escherichia coli*. Producer strains leak out amino acids which are costly to produce, and can be taken up by mutants lacking the ability to synthesize them. Several different amino acids were investigated, with up to two at a time spanning

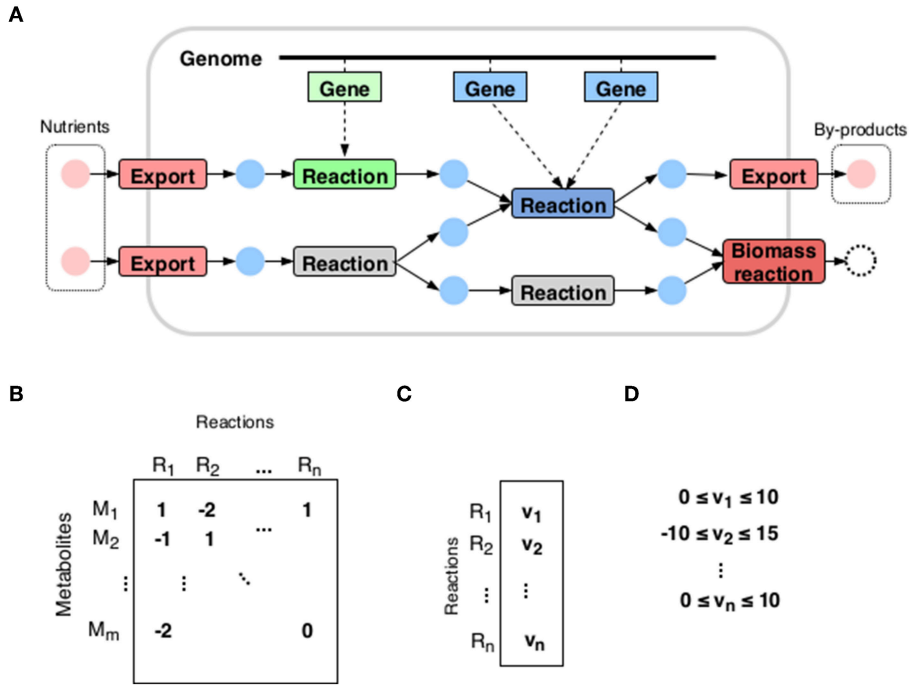


FIGURE 1 | A metabolic reconstruction. **(A)** A *metabolic network* is constructed by joining together reactions whose enzymes are recognized in the organism's genome. **(B)** The most commonly used mathematical representation is the *stoichiometric matrix*: its rows correspond to the metabolites of the network and columns to reactions. The entries are stoichiometric coefficients: how much of the metabolite is consumed or produced in the reaction. **(C)** The *flux vector* expresses the state of the metabolism: its entries correspond to reaction rates. **(D)** Each reaction rate has a lower and an upper bound.

four possible strategies (genotypes). Again, both the level of leakiness and the cost of production influence the type of equilibria observed. With low enough levels of leakiness, both an equilibrium with a full producer coexisting with a complete auxotroph, as well as cross-feeding are possible. With increasing leakiness, the full producer becomes non-viable. However, it was also observed that due to interdepencies in amino acid production, in some situations cross-feeding is not possible because losing the ability to produce one amino acid leads to the loss of the ability to produce the other. Zomorrodi and Segrè also studied the evolutionary dynamics of these interactions by performing *in silico* invasion experiments. They found that cross-feeding can emerge through the progressive loss of amino acid synthesis capabilities, and that this mutually dependent coalition is often stable against invasion by non-producers, consistent with previous experimental findings [31, 32].

2. YIELD VS. RATE

One of the questions already extensively explored through applying game theory to metabolism is ATP production. There is a fundamental trade-off between yield and rate of ATP production in heterotrophic organisms: some of the free energy obtained from substrate degradation is needed to drive the reaction. Increasing the portion of free energy that is used for driving the reaction increases the rate of ATP production but lowers the yield. The choice of pathway thus presents a social dilemma. Choosing the efficient strategy would maximize resource usage and benefit the population as a whole. However, if an individual cell chooses to stray from this cooperative path, its faster growth rate will allow it to increase in numbers and eventually overcome the cooperators at the cost of the interest of the community.

In Pfeiffer et al. [33], this question is explored in the context of respiration vs. fermentation. This paper is to our knowledge the first to apply game theory specifically to metabolic pathways. Most organisms can in principle choose to degrade sugar by both the respiration and fermentation pathways. While fermentation provides ATP faster, it has a significantly lower yield. Thus, fermentation can be seen as a wasteful, "selfish" strategy, while respiration is more efficient in terms of nutrient use.

By constructing a simple population model, the authors show that while a population of fermenters will be smaller due to a faster depletion of resources, they can nevertheless take over a population of respirators due to their faster growth rate. This constitutes the famous "tragedy of the commons" [34]. However, if a spatial component is added, respirators can have a chance. This is because at lower nutrient levels, fermenters will deplete their immediate environment of resources and suffer the consequences.

Frick and Schuster [35] explore this question further. They too construct a population model for slow but efficient vs. fast but wasteful resource use. The authors then interpret the steady state population densities of both strategies in each different scenario

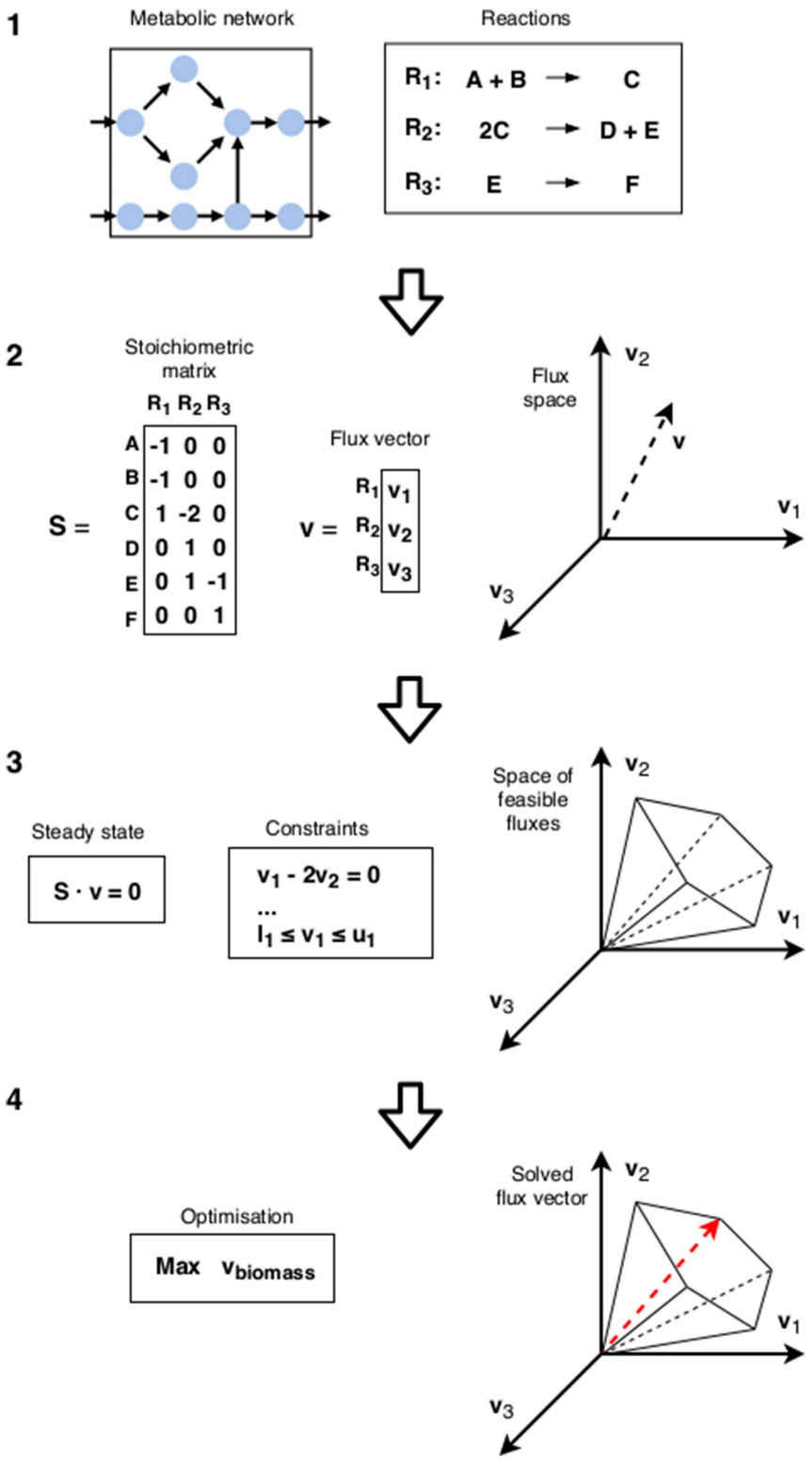


FIGURE 2 | Flux Balance Analysis: **(1)** The starting point is the metabolic reconstruction. **(2)** The *flux space* is a space of possible flux vectors v. **(3)** Constraints are imposed on the flux space: the *steady state* condition requires that the net production of all internal metabolites equals zero. This confines the flux vector into a cone-shaped subspace. The upper and lower bounds (l_i and u_i) of the reactions bound the cone, establishing a maximum magnitude. **(4)** Finally, using linear programming, an objective function (usually a linear combination of certain reaction fluxes that corresponds to a biological objective such as biomass) is maximized to find the predicted flux vector.

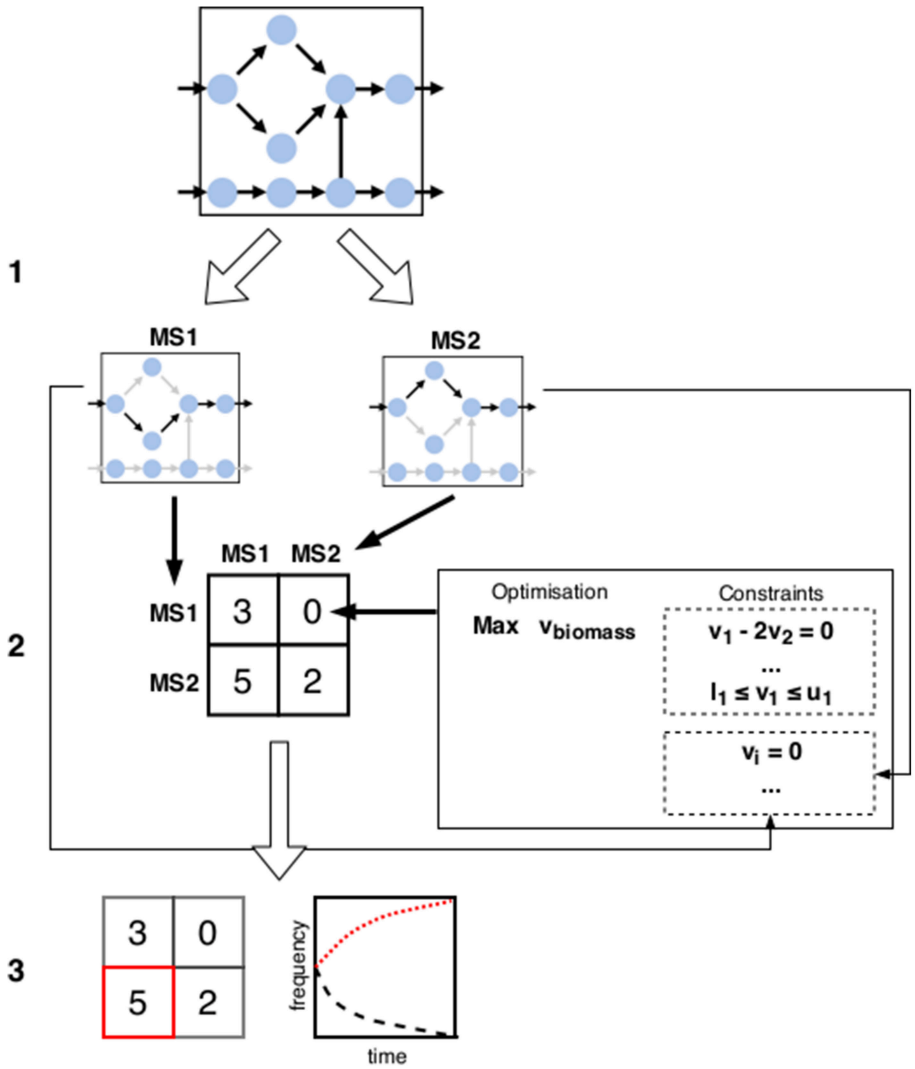


FIGURE 3 | Schematic overview of the tentative framework for metabolic games: **(1)** Starting from the metabolic network(s) of one or several organisms, possible metabolic phenotypes are subdivided into distinct *metabolic strategies* (denoted by MS1 and MS2 above). **(2)** *Payoffs* for each configuration of strategies are calculated using a method such as FBA. New constraints are added in each scenario to reflect the choice of phenotype implicated by the strategy as well as possible environmental interactions due to the strategy choices of other players (e.g., changes in substrate availability can be enforced by restricting fluxes of import reactions accordingly). **(3)** Predicted configuration(s) of metabolic strategies is determined by looking for the solutions of the game (e.g., finding the Nash equilibria or studying the dynamics of the strategies).

as payoffs: in this way, the situation is a Prisoner's Dilemma with pure respiration forming the cooperation strategy. This is important because were the growth rates to be taken as the payoffs, one would conclude that fermentation is the optimal choice in all instances. However, from the point of view of sustaining the highest possible population density, cooperation, that is respiration, is the best choice.

Experimental evidence for the results described above was provided in MacLean and Gudelj [36]. The authors used yeast as their model organism and grew pure respirators and respiro-fermenters together in different culture set-ups. They found that while the "cheaters" win in a chemostat, in serial batch and spatially structured populations, the two strategies can coexist.

Schuster et al. [10] critically examined the assumption made in FBA of maximization of biomass yield. They argued that in general there is a trade-off between yield and rate, and that it is not *a priori* clear which of these conflicting goals would be selected for. Based on the theoretical results previously put forth by Pfeiffer et al. [33] as well as several examples from nature, the authors conclude that maximization of yield cannot be considered a universal principle.

Aledo et al. [37] also studied the yield vs. rate question but this time in glycolysis itself, which can operate under two different regimes: one with a high yield but a slower rate, another with a low yield but a faster rate. Using a simple matrix game model, with payoffs derived as functions of extracellular free energy and

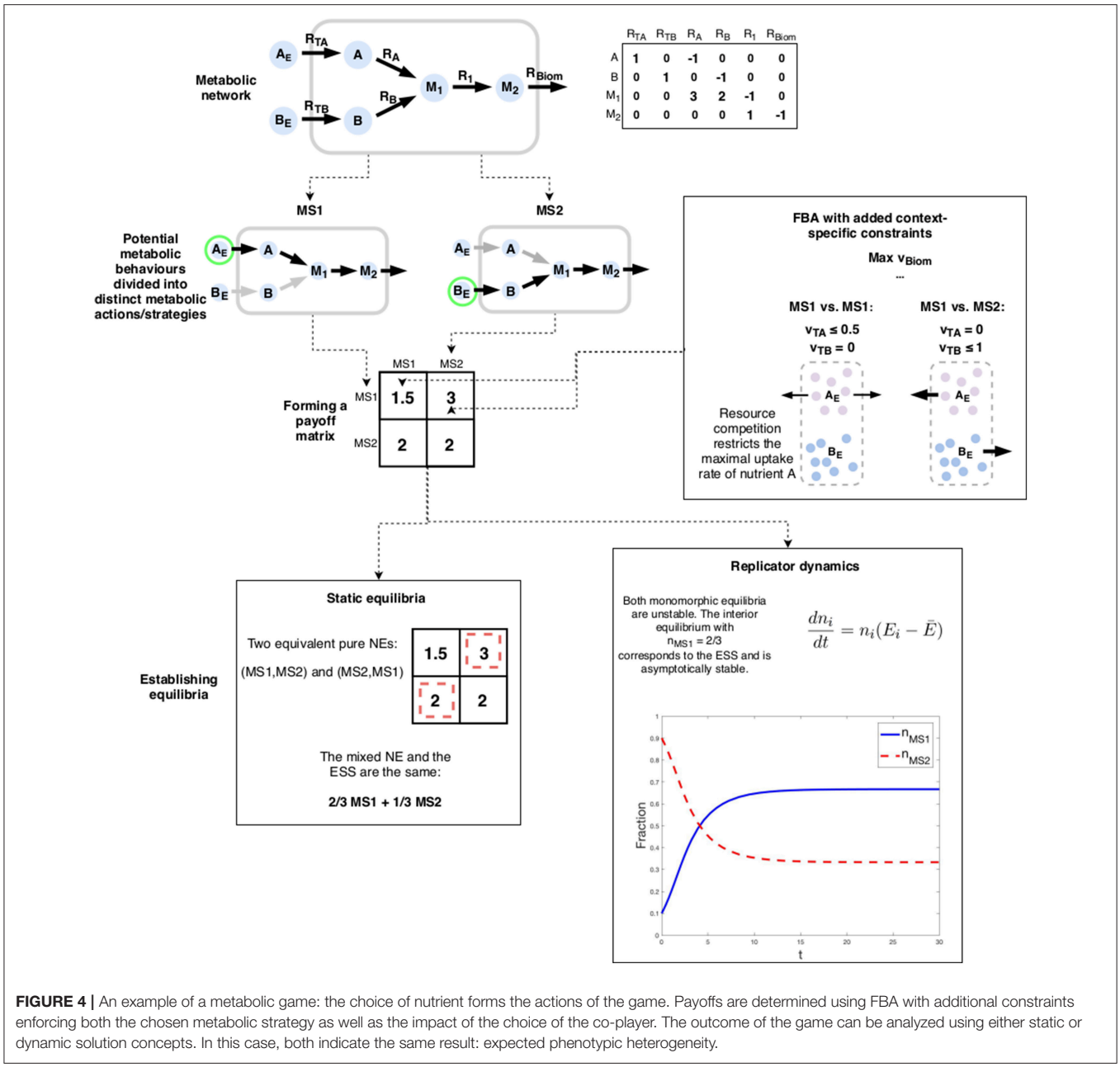


FIGURE 4 | An example of a metabolic game: the choice of nutrient forms the actions of the game. Payoffs are determined using FBA with additional constraints enforcing both the chosen metabolic strategy as well as the impact of the choice of the co-player. The outcome of the game can be analyzed using either static or dynamic solution concepts. In this case, both indicate the same result: expected phenotypic heterogeneity.

in agreement with the Prisoner's Dilemma payoff scheme, the authors showed that in a well-mixed population, cooperation cannot persist. In contrast, if the game is played on a lattice so that players only interact with their neighbors, cooperation is a possible outcome.

Schuster et al. [38] returned to the question of yield vs. rate. They presented a toy model representing a simplified version of ATP production to show that whether maximizing the yield coincides with maximizing the rate depends on the particulars of the system. They also further articulated the idea that alternative pathways can be seen as strategies in the game theoretical sense, and that "choosing" which pathway to use can happen not only through changes in genotype, but also through regulatory changes within the life-span of a cell.

Kareva [39] investigated the yield vs. rate question in the context of cancer cells where the use of the more inefficient glycolysis pathway is observed as one of the hallmarks of cancerous growth and is known as the Warburg effect [40, 41]. However, in contrast to the previous models, the author argued that the use of glycolysis is the cooperative strategy: while recognizing the possibility to increase the rate of glucose uptake, she considered the use of glycolysis to remain detrimental to the individual cell due to its low yield. Meanwhile, the associated lactic acid production can benefit the cancer cell population

as a whole, if undertaken in sufficient numbers, because it disproportionately harms non-cancerous cells. Thus, glycolysis can be considered as public goods production. The contradiction with previous studies is clear. However, in the ODE system used to model a population of cells with varying rates of carbon allocated to glycolysis in Kareva [39], it was observed that glycolytic cells do increase in frequency if they have a faster growth rate.

In two successive papers [42, 43], Archetti presented a public goods model of the Warburg effect. He took the same view as Kareva [39] and considered glycolysis as the cooperative strategy amongst cancer cells. The benefit accrued by all participants from glycolysis—increased acidity—is modeled by a double sigmoid function: increased acidity yields a benefit over healthy cells if enough cells are producing lactic acid, but too much will start to hamper the growth of even cancer cells. The dynamics of the frequencies of glycolytic and non-glycolytic cells were modeled using the replicator equation. Because an exact solution of the dynamics for a sigmoid shaped benefit is not available, Bernstein polynomials were used to find an approximate solution. Archetti found that if the cost attached to glycolysis is not too high, glycolytic cooperators can persist at intermediate frequencies.

3. PUBLIC GOODS

Another possible social dilemma within microbial communities occurs with necessary but costly functions. If a metabolic function is performed at the cell surface or outside the cell, it means that the benefit incurred can be shared by other cells that are possibly not contributing to the undertaking of the said function. Such a situation is best described by a public goods game.

Gore et al. [44] studied the invertase production system of *S. cerevisiae* described in section . Their model is a sort of mix between a public goods game and a matrix game: the authors define payoffs in terms of the fraction of invertase-producers in the population but then go on to compare these payoff values to the well-known 2-player games. If the benefits are linear, cooperation cannot persist unless the benefit derived from sucrose degradation by the invertase-producer exceeds the cost, in which case producing the enzyme is not a public good. On the other hand, with non-linear benefits, frequency-dependent selection allows for a fraction of the cooperators to persist. This result was in line with experimental evidence which confirmed both the coexistence of producers and non-producers as well as the non-linear benefit function.

A similar model was presented in Schuster et al.[45]. In this paper, Schuster et al. studied generic exoenzyme production assuming again that some fraction of the transformed growth product diffuses directly into the producer cell while the rest is available to the surrounding community. This time the benefit from the public good is given by a Monod function modeling the growth rate attained through the available nutrient. The nutrient acquired in turn depends on the fraction of cooperators in the population and cell density, which is a parameter of the model. The authors conclude that depending on the parameters, the fraction of public good that diffuses away, the cost of enzyme production, and the cell density, the model can be seen as a Prisoner's Dilemma, a Snowdrift or a Harmony game.

Archetti [46] studied growth factor production in cancer cells as a public goods game. Growth factor production is costly but the benefits are available to all surrounding cells. The benefit function was assumed to have sigmoid shape and population dynamics were modeled by the replicator equation. As in Archetti [42, 43], Bernstein polynomials were used to circumvent the problem caused by the sigmoid function. Archetti found that depending on how exactly the fraction of producers influences the benefit from growth factor, different types of dynamics are possible: a globally attracting mixed equilibrium where producers and non-producers coexist, the fixation of one type depending on the initial frequencies, or the fixation of producers regardless of the initial conditions.

The model presented in Archetti [46] was expanded on in Archetti [47] by introducing a spatial component. In this model, cells are placed in the nodes of a Voronoi graph. A Voronoi graph has the average connectivity of 6, with very few nodes beyond degree 4–8. Cells receive benefits from growth factors produced by producer-cells within a neighborhood defined by a diffusion parameter, discounted with the distance to the focal cell. The benefit itself is given by a normalized logistic function. In other words, benefits are non-linear. Archetti found that similar to well-mixed populations, cooperation declines as the cost of production increases. Stochasticity in the update rules used to model proliferation and a steeper benefit function also decrease cooperation.

4. NUTRIENT CHOICE

Perhaps the best examples showcasing the usefulness of game theoretic thinking are situations where frequency-dependent selection leads to polymorphisms in nutrient use. It is often the case that in a given environment, there is a preferred choice for the main carbon source. However, in any realistic scenario, nutrient availability is limited, and it can be beneficial for the individual to opt for a carbon source that is slightly less optimal, but abundant due to being the "unpopular" choice.

Doebeli [48] considered the evolution of cross-feeding. He constructed a model for a bacterial culture growing in a chemostat, using glucose as its main nutrient. During growth on glucose, acetate is secreted which can also be used as a nutrient, albeit with a lower growth rate. Doebeli assumed that there is a trade-off in using the secondary metabolite: becoming more proficient in using acetate lowers the ability to use glucose efficiently. Furthermore, this trade-off is subject to gradual change through mutations. Bacterial growth and nutrient concentration was modeled using a Michaelis-Menten type model.

Using the theory of adaptive dynamics, Doebeli showed that the frequency-dependent selection following from the trade-off can lead to evolutionary branching and the emergence of a stable polymorphism of glucose and acetate specialists. He also found that if the dynamics are changed to model a serial batch culture

instead of a chemostat, evolution of cross-feeding becomes much less likely. In a chemostat culture, the concentration of nutrients is kept constant, while in a batch culture nutrients are allowed to be depleted. These results were further expanded and provided experimental confirmation in Friesen et al. [49].

Kianercy et al. [50] studied the Warburg effect and the *reverse* Warburg effect. The reverse Warburg effect refers to the phenomenon wherein some cells in a tumor use lactate secreted as a by-product of glycolysis as their energy source. The authors' model is a 2-player matrix game with two types of players: hypoxic and oxygenated cells. Both types have the same available strategies: using either glucose or lactate as their nutrient. Lactate is secreted by hypoxic cells using glucose. Similarly to Kareva [39] and Archetti [42, 43], the authors take yields as payoffs. Thus, using glucose gives a lower payoff for hypoxic cells. The authors found that there exist two stable states and conclude that lactate secretion can induce a transition between high and low levels of glucose consumption.

Healey et al. [51] investigated phenotypic *bet-hedging* by experiments and a game theory model. Bet-hedging refers to a hypothesis that microbes may increase their survival in fluctuating environments by implementing a stochastic phenotype. In other words, a genetically homogeneous population might display two (or more) distinct phenotypes. In the language of game theory, this would constitute a mixed strategy. The model system in Healey et al.[51] was *S. cerevisiae* that prefers glucose as its carbon source, but also harbors the GAL network for metabolizing galactose. The game theory model used was a simple foraging game, where a population of players must choose between two resources. One of the resources is the preferred one, and so there is an additional cost associated with using the inferior resource. However, if all members of the population have chosen the preferred resource, it is better for an individual to choose the other. This leads to a stable mixed equilibrium of users of both resources. Experiments performed by Healey et al. corroborated this theoretical result.

5. DISCUSSION

In this paper, we presented the idea of a metabolic game and reviewed the main existing literature on applying evolutionary game theory to the study of metabolism. Most studies so far have evoked game theory as an explanatory device, making use of established knowledge on famous games such as the Prisoner's Dilemma to qualitatively describe specific observed phenomena. We believe that it is possible to go beyond that, and to develop a formalism for the metabolic modeling of multicellular and multispecies communities by combining the ideas behind evolutionary game theory with the existing tools of constraint-based modeling.

The recent paper by Zomorrodi and Segre [23] is a first step in this direction. However, there are significant challenges remaining before the game theoretical perspective can be taken full advantage of. Namely, properly defining the different components of a game must be carefully considered in order to make the models derived as reliable and descriptive as possible.

The first component of a game are the players. They are the participants in the interaction under study. Many of the papers discussed in this survey used some form of a 2-player matrix game to make their arguments. In principle, this type of game represents a situation where two individuals face each other in a single interaction to obtain a single payoff. With this in mind, it seems strange to use this model when talking about microbial populations. However, when the matrix game is embedded in the replicator dynamics or another kind of frequency dependent selection model, it starts to more closely resemble a microbe culture. In a way, payoffs are obtained according to who one's average neighbor is at any given time, as might be imagined to happen in a well-mixed culture.

Yet the question remains if pairwise encounters are sufficient to capture the interaction dynamics of microbes that mainly influence each other through diffusible molecules. The other type of model often used is the public goods game. At first glance it seems to more accurately describe an interaction through diffusible molecules because it considers several players to take part in the game at the same time. For example, in the case of invertase production, it is intuitive to consider the game to comprise those cells that the released glucose can be assumed to reach. However, there are some problems with using the public goods game as a general model. Firstly, the benefit function must be accurately estimated since its form can greatly influence the type of dynamics it gives rise to (see for example [46], see also [44]). This might be difficult to do without experimental evidence. Secondly, public goods games with non-linear benefit functions can be difficult to analyse [46], although some progress has been made in this area recently [52].

Explicit consideration of spatial structure could facilitate properly defining interacting agents. Even if the underlying model is a 2-player game, embedding it into a spatial model so that individuals interact with those around them, and the changes resulting in the environment from these actions (depletion of nutrients etc.) happen locally, will be more faithful to nature. The standard way to represent spatial structure in game theory is to assign players to nodes in a graph as was done in Archetti [47]. This approach might be most applicable to environments such as biofilms. The other option is to use partial differential equations to include spatial dimensions in the population dynamics. The main problem with both approaches is that usually the only analysis possible is through simulations. Furthermore, parameters such as diffusion coefficients might be needed to specify the model.

Considering all of the above, it seems that if the goal is to specify a systematic framework in which a metabolic game can be defined based mainly on the metabolic reconstructions of the organisms, the simple matrix game should be the model of choice. Indeed, in order to have a computational framework anywhere close to the simplicity of the original FBA formalism, it seems that only the high level ideas from game theory, mainly considering the choice available for one individual in conjunction with the choices available to their opponents, can be included. This is already captured by the matrix game. In addition, authors have arrived at similar conclusions modeling the same situation with

various more complicated models [44, 45] and the simpler matrix game [23].

With regard to the choice of action/strategy space, the question is mostly a technical one. In principle, a game constructed on the basis of metabolic networks would consider as available actions the range of feasible metabolic behaviors, in other words, the flux cone [53, 54]. However, from a practical standpoint it is evident that some abstraction is needed. Firstly, in order for a matrix game to be defined, the action space needs to be discretized. Secondly, with the number of reactions and thus of flux values to be defined routinely reaching to thousands, the game would surely quickly become intractable.

Several approaches to a decomposition of the flux cone have been proposed. Most notably, three related concepts, *elementary flux modes* [55, 56], *extreme currents* [53], and *extreme pathways* [57] all formulate a mathematical definition of a pathway using concepts from linear algebra and convex analysis. Using such concepts, the space of available metabolic phenotypes can be characterized in terms of which reactions are active, each set corresponding roughly to separate biochemical pathways that are able to operate at a steady state. Unfortunately, the number of elements in such a decomposition grows exponentially with the size of the network [58–60]. It might thus be impossible in practice to define the action space simply using these concepts, at least at the level of genome-scale reconstructions. De Figueiredo et al. [61] have offered a possible amendment by proposing an efficient procedure to compute elementary flux modes in order of increasing number of reactions.

Other concepts worth exploring are the *phenotypic phase plane* put forth by Edwards et al. [62] and the *flux tope* by Gerstl et al. [63]. A phenotypic phase plane is defined by the uptake rates of two nutrients. The optimal metabolic behavior is calculated at each point of the plane using a biomass function. It turns out that such a plane is divided into a finite number of distinct regions with qualitatively different metabolic behavior. A flux tope is obtained by specifying a direction for all reversible reactions. It corresponds to a maximal "pathway" (as opposed to a minimal one, such as an elementary flux mode). The authors report that the calculation of all flux topes is possible even at a larger scale.

In Zomorrodi and Segre [23], available metabolic actions were not defined explicitly in terms of flux distributions but rather by excreted compounds. One or several metabolites of interest were first forced to be exported and hence produced (or alternatively to *not* be produced simulating auxotrophy), after which the metabolic state can be determined using standard optimization principles with the additional constraints. There are compelling arguments for defining actions in metabolic games using extracellular compounds. In general, microbial interactions are often mediated by the exchange of molecules. By focusing on these compounds, the elements of the action space have a clear interpretation in the context of interaction. The set of possible secretions is also much more tractable than the space of all possible metabolic phenotypes.

Interactions based on extracellular metabolites were characterized from a slightly different point of view in Klitgord and Segre [64]. The authors asked whether it is possible to predict species interactions based on culture media. Using genome-scale stoichiometric models they tested whether growth of two organisms was possible in isolation and in tandem in a given medium. This approach showed examples of both mutualistic and commensal relationship induced by growth media.

Wintermute and Silver [65] used a similar model to study the exchange of metabolites. The authors showed how the costs and benefits of extracellular metabolites can be estimated using the concept of shadow prices from constraint-based analysis. The shadow price of a metabolite can be understood as a measure of how much the objective, for example biomass flux, would change if the production of the said metabolite changed. Such an analysis could be very useful for metabolic games since it allows one to compute both the cost of producing a diffusible molecule as well as the benefit derived from it by the organism that is able to receive it.

In a thesis work, Wannagat [66] showed how to compute the minimal sets of compounds two organisms need to exchange in order to be able to grow. Here the approach was qualitative and was used to categorize interactions in terms of their type, but such a procedure could be used also to define the action space in a metabolic game.

Finally, in order to construct a game, one needs to define the payoffs. This is arguably the most crucial step since the payoff values will largely determine the predictions of the model. There is a particular importance to not only qualitatively, but also quantitatively establish accurate payoffs here since the hope is for metabolic game theory to match the predictive ability of FBA. One example from the literature discussed in this paper highlights both the importance and the difficulty in defining payoffs.

In several papers [10, 33, 35, 36], fermentation in the presence of oxygen is seen as a classic "cheater" strategy. From an individual's point of view, the inefficiency of fermentation in terms of yield is not "seen": what the cell experiences as the consequence of its choice is a growth rate exceeding that of its conspecifics. The result of a wasteful use of resources is only felt at the population level, resulting in a lower sustainable cell density. This is the (in)famous Prisoner's Dilemma. However, when essentially the same situation has been discussed in the context of cancer [39, 42, 47], a completely opposite view has been adopted. Here, fermentation was seen as the cooperation strategy. For example, Archetti [42] described using fermentation as a contribution to a public good, the cost of the action being the loss in yield compared to respiration. While it can be argued that the underlying biology is very different for single-celled microbes and cancerous tissue, the discrepancy is still puzzling.

The problem of properly defining payoffs in the yield vs. rate dilemma is related to that of normalization in FBA [10]. In order to "ground" the flux vector, normalization is needed. A common choice for a numeraire is the uptake of a primary nutrient. The fact that maximization of flux through the biomass reaction in FBA leads to a *de facto* maximization of biomass yield follows from this operation. Consider now the situation in ATP production. If the value of the objective function in a standard FBA approach is taken as the payoff, respiration is a better strategy than fermentation. However, as already discussed, a fermenter can outgrow its respiring neighbor. From

the perspective of evolutionary game theory, it is thus clearly the winner, and its payoff should reflect this fact. However, if we simply switch the payoff from yield to actual rate of biomass production, two fermenters would also obtain the highest payoff together. This is because we have assumed in a simplified way that the external resources are infinite, and hence two fermenters are able to sustain the increased uptake of nutrients they achieve in the presence of respirators. In order to arrive at the Prisoner's Dilemma payoff structure, we need to take into account that if everyone uses fermentation, it can no longer provide the benefit it has over respiration because of a depletion of nutrients.

The above example showcases the difficulty in appropriately quantifying the outcomes in a metabolic game. Optimization of an appropriate objective function can certainly accurately identify "catastrophic" outcomes where growth is not possible, but when conclusions are drawn as to which metabolic strategy would win in intra- or interspecific competition, caution is warranted. One must make sure that the quantity under consideration is apt to decide the winner(s) in an evolutionary sense.

The definition of the action space can also offer a way to quantify the payoffs. For example, if different metabolic phenotypes are characterized by imported and exported metabolites, benefits and costs can be calculated following [65]. This could open the way for a more systematic definition of public goods games using only the knowledge obtained from metabolic models.

To further develop these ideas, finding new suitable model organisms, especially for interspecies interactions, would be of great interest. However, if a generally applicable framework for metabolic games is desired, it is important to avoid overfitting the model to specific situations. Since for single-species communities the work of Zomorrodi and Segre [23] already offers an excellent starting point, a good goal for future research would be a systematic definition of the action space that does not rely on context-specific biological information. With regard to multispecies interactions, this area of research remains less explored. Thus, even a proof of concept application with metabolic strategies derived based on biological knowledge would be desirable.

Besides games, other models from economics have generated interest in the field of microbiology. The concept of comparative advantage [67] was thus applied to gene circuits in Enyeart et al. [68]. The authors showed that when two bacterial species trade signaling molecules necessary for survival, they both enjoy improved growth, as predicted by the theory of comparative advantage. Tasoff et al. [69] used general equilibrium theory [70] to understand the mutualistic exchange of compounds between micro-organisms. The authors argued that comparative advantage is a necessary condition for the exchange to take place. This theory can be further extended to several organisms exchanging multiple compounds. Other concepts that have been suggested for applications in the microbial context include avoidance of bad trading partners, establishment of local business ties, diversification or specialization, monopolization of a market, and elimination of competitors [71].

AUTHOR CONTRIBUTIONS

TP, MW, and M-FS conceived the idea. TP and MW performed the literature search. TP wrote the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This study was funded by the Horizon 2020 Program of the European Commission within the Marie Skłodowska-Curie Innovative Training Network MicroWine (grant number 643063).

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

REFERENCES

1. Varma A, Palsson BO. Metabolic flux balancing: basic concepts, scientific and practical use. *Biotechnology*. (1994) **12**:994–8. doi: 10.1038/nbt1094-994
2. Edwards JS, Palsson BO. Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*. (2000) **1**:1. doi: 10.1186/1471-2105-1-1
3. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol*. (2001) **19**:125–30. doi: 10.1038/84379
4. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol*. (2010) **28**:245–8. doi: 10.1038/nbt.1614
5. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell*. (2015) **161**:971–87. doi: 10.1016/j.cell.2015.05.019
6. Gottstein W, Olivier BG, Bruggeman FJ, Teusink B. Constraint-based stoichiometric modelling from single organisms to microbial communities. *J R Soc Interface*. (2016) **13**:20160627. doi: 10.1098/rsif.2016.0627
7. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, et al. Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol*. (2007) **3**:92. doi: 10.1038/msb4100131
8. Zomorrodi AR, Maranas CD. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol*. (2012) **8**:e1002363. doi: 10.1371/journal.pcbi.1002363
9. Khandelwal RA, Olivier BG, Röling WF, Teusink B, Bruggeman FJ. Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE*. (2013) **8**:e64567. doi: 10.1371/journal.pone.0064567
10. Schuster S, Pfeiffer T, Fell DA. Is maximization of molar yield in metabolic networks favoured by evolution? *J Theor Biol*. (2008) **252**:497–504. doi: 10.1016/j.jtbi.2007.12.008
11. Schuster S, Kreft JU, Schroeter A, Pfeiffer T. Use of game-theoretical methods in biochemistry and biophysics. *J Biol Phys*. (2008) **34**:1–17. doi: 10.1007/s10867-008-9101-4
12. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, et al. Improving metabolic flux predictions using absolute gene expression data. *BMC Syst Biol*. (2012) **6**:73. doi: 10.1186/1752-0509-6-73

13. Von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press (2007).
14. Smith JM, Price G. The logic of animal conflict. *Nature*. (1973) **246**:15–8. doi: 10.1038/246015a0
15. Smith JM. *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press (1982).
16. Pfeiffer T, Schuster S. Game-theoretical approaches to studying the evolution of biochemical systems. *Trends Biochem Sci*. (2005) **30**:20–5. doi: 10.1016/j.tibs.2004.11.006
17. Ruppin E, Papin JA, De Figueiredo LF, Schuster S. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr Opin Niotechnol*. (2010) **21**:502–10. doi: 10.1016/j.copbio.2010.07.002
18. Hummert S, Bohl K, Basanta D, Deutsch A, Werner S, Theißen G, et al. Evolutionary game theory: cells as players. *Mol Biosyst*. (2014) **10**:3044–65. doi: 10.1039/C3MB70602H
19. Nash JF, Jr. Equilibrium points in n-person games. *Proc Natl Acad Sci USA*. (1950) **36**:48–9. doi: 10.1073/pnas.36.1.48
20. Nash J. Non-cooperative games. *Ann Math*. (1951) **54**:286–95. doi: 10.2307/1969529
21. Haigh J. Game theory and evolution. *Adv Appl Probab*. (1975) **7**:8–11. doi: 10.2307/1425844
22. Avis D, Rosenberg GD, Savani R, Von Stengel B. Enumeration of Nash equilibria for two-player games. *Econ Theor*. (2010) **42**:9–37. doi: 10.1007/s00199-009-0449-x
23. Zomorrodi AR, Segrè D. Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nat Commun*. (2017) **8**:1563. doi: 10.1038/s41467-017-01407-5
24. Broom M, Rychtář J. *Game-Theoretical Models in Biology*. Boca Raton, FL: CRC Press (2013).
25. Archetti M, Scheuring I. Review: game theory of public goods in one-shot social dilemmas without assortment. *J Theor Biol*. (2012) **299**:9–20. doi: 10.1016/j.jtbi.2011.06.018
26. Hofbauer J, Sigmund K. *Evolutionary Games and Population Dynamics*. Cambridge, UK: Cambridge University Press (1998).
27. Page KM, Nowak MA. Unifying evolutionary dynamics. *J Theor Biol*. (2002) **219**:93–8. doi: 10.1016/S0022-5193(02)93112-7
28. Metz JA, Geritz SA, Meszéna G, Jacobs FJ, Van Heerwaarden JS. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. *Stochast Spatial Struct Dynam Syst*. (1996) **45**:183–231.
29. Dieckmann U, Law R. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *J Math Biol*. (1996) **34**:579–612. doi: 10.1007/BF02409751
30. Geritz SAH, Kisdi E, Meszena G, Metz JAJ. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evol Ecol*. (1998) **12**:35–57. doi: 10.1023/A:1006554906681
31. Pande S, Merker H, Bohl K, Reichelt M, Schuster S, De Figueiredo LF, et al. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J*. (2014) **8**:953. doi: 10.1038/ismej.2013.211
32. Campbell K, Vowinckel J, Mülleder M, Malmsheimer S, Lawrence N, Calvani E, et al. Self-establishing communities enable cooperative metabolite exchange in a eukaryote. *Elife*. (2015) **4**:e09943. doi: 10.7554/eLife.09943
33. Pfeiffer T, Schuster S, Bonhoeffer S. Cooperation and competition in the evolution of ATP-producing pathways. *Science*. (2001) **292**:504–7. doi: 10.1126/science.1058079
34. Lloyd WF. *Two Lectures on the Checks to Population: Delivered Before the University of Oxford, in Michaelmas Term 1832*. Oxford, UK: JH Parker (1833).
35. Frick T, Schuster S. An example of the prisoner's dilemma in biochemistry. *Naturwissenschaften*. (2003) **90**:327–31. doi: 10.1007/s00114-003-0434-3
36. MacLean RC, Gudelj I. Resource competition and social conflict in experimental populations of yeast. *Nature*. (2006) **441**:498. doi: 10.1038/nature04624
37. Aledo JC, Pérez-Claros JA, Del Valle AE. Switching between cooperation and competition in the use of extracellular glucose. *J Mol Evol*. (2007) **65**:328–39. doi: 10.1007/s00239-007-9014-z
38. Schuster S, de Figueiredo LF, Schroeter A, Kaleta C. Combining metabolic pathway analysis with evolutionary game theory. Explaining the occurrence of low-yield pathways by an analytic optimization approach. *Biosystems*. (2011) **105**:147–53. doi: 10.1016/j.biosystems.2011.05.007
39. Kareva I. Prisoner's dilemma in cancer metabolism. *PLoS ONE*. (2011) **6**:e28576. doi: 10.1371/journal.pone.0028576
40. Warburg O, Wind F, Negelein E. Ueber den stoffwechsel von tumoren im körper. *Klinische Wochenschrift*. (1926) **5**:829–32. doi: 10.1007/BF01726240
41. Warburg O. On the origin of cancer cells. *Science*. (1956) **123**:309–14. doi: 10.1126/science.123.3191.309
42. Archetti M. Evolutionary dynamics of the Warburg effect: glycolysis as a collective action problem among cancer cells. *J Theor Biol*. (2014) **341**:1–8. doi: 10.1016/j.jtbi.2013.09.017
43. Archetti M. Heterogeneity and proliferation of invasive cancer subclones in game theory models of the Warburg effect. *Cell Prolif*. (2015) **48**:259–69. doi: 10.1111/cpr.12169
44. Gore J, Youk H, Van Oudenaarden A. Snowdrift game dynamics and facultative cheating in yeast. *Nature*. (2009) **459**:253–6. doi: 10.1038/nature07921
45. Schuster S, Kreft JU, Brenner N, Wessely F, Theißen G, Ruppin E, et al. Cooperation and cheating in microbial exoenzyme production–theoretical analysis for biotechnological applications. *Biotechnol J*. (2010) **5**:751–8. doi: 10.1002/biot.200900303
46. Archetti M. Evolutionary game theory of growth factor production: implications for tumour heterogeneity and resistance to therapies. *Br J Cancer*. (2013) **109**:1056–62. doi: 10.1038/bjc.2013.336
47. Archetti M. Cooperation among cancer cells as public goods games on Voronoi networks. *J Theor Biol*. (2016) **396**:191–203. doi: 10.1016/j.jtbi.2016.02.027
48. Doebeli M. A model for the evolutionary dynamics of cross-feeding polymorphisms in microorganisms. *Popul Ecol*. (2002) **44**:59–70. doi: 10.1007/s101440200008
49. Friesen ML, Saxer G, Travisano M, Doebeli M. Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in *Escherichia coli*. *Evolution*. (2004) **58**:245–60. doi: 10.1111/j.0014-3820.2004.tb01642.x
50. Kianercy A, Veltri R, Pienta KJ. Critical transitions in a game theoretic model of tumour metabolism. *Interface Focus*. (2014) **4**:20140014. doi: 10.1098/rsfs.2014.0014
51. Healey D, Axelrod K, Gore J. Negative frequency-dependent interactions can underlie phenotypic heterogeneity in a clonal microbial population. *Mol Syst Biol*. (2016) **12**:877. doi: 10.15252/msb.20167033
52. Archetti M. How to analyze models of nonlinear public goods. *Games*. (2018) **9**:17. doi: 10.3390/g9020017
53. Clarke BL. Stability of complex reaction networks. *Adv Chem Phys*. (1980) **43:** 1–215. doi: 10.1002/9780470142622.ch1
54. Clarke BL. Stoichiometric network analysis. *Cell Biophys*. (1988) **12**:237–53. doi: 10.1007/BF02918360
55. Schuster R, Schuster S. Refined algorithm and computer program for calculating all non–negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Bioinformatics*. (1993) **9**:79–85. doi: 10.1093/bioinformatics/9.1.79
56. Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. *J Biol Syst*. (1994) **2**:165–82. doi: 10.1142/S0218339094000131
57. Schilling CH, Edwards JS, Letscher D, Palsson BØ. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng*. (2000) **71**:286–306. doi: 10.1002/1097-0290(2000)71:4<286::AID-BIT1018>3.0.CO;2-R
58. Klamt S, Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*. (2002) **29**:233–6. doi: 10.1023/A:1020390132244
59. Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandekar T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics*. (2002) **18**:351–61. doi: 10.1093/bioinformatics/18.2.351
60. Acuna V, Chierichetti F, Lacroix V, Marchetti-Spaccamela A, Sagot MF, Stougie L. Modes and cuts in metabolic networks: complexity and algorithms. *Biosystems*. (2009) **95**:51–60. doi: 10.1016/j.biosystems.2008.06.015
61. De Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, et al. Computing the shortest elementary flux modes in

genome-scale metabolic networks. *Bioinformatics*. (2009) **25**:3158–65. doi: 10.1093/bioinformatics/btp564
62. Edwards JS, Ramakrishna R, Palsson BO. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng*. (2002) **77**:27–36. doi: 10.1002/bit.10047
63. Gerstl MP, Müller S, Regensburger G, Zanghellini J. Flux tope analysis: studying the coordination of reaction directions in metabolic networks. *Bioinformatics*. (2018) **35**:266–73. doi: 10.1093/bioinformatics/bty550
64. Klitgord N, Segrè D. Environments that induce synthetic microbial ecosystems. *PLoS Comput Biol*. (2010) **6**:e1001002. doi: 10.1371/journal.pcbi.1001002
65. Wintermute EH, Silver PA. Emergent cooperation in microbial metabolism. *Mol Syst Biol*. (2010) **6**:407. doi: 10.1038/msb.2010.66
66. Wannagat M. *Study of the Evolution of Symbiosis at the Metabolic Level Using Models from Game Theory and Economics*. Université de Lyon (2016).
67. Ricardo . *On the Principles of Political Economy and Taxation*. London: John Murray (1817).
68. Enyeart PJ, Simpson ZB, Ellington AD. A microbial model of economic trading and comparative advantage. *J Theor Biol*. (2015) **364**:326–43. doi: 10.1016/j.jtbi.2014.09.030
69. Tasoff J, Mee MT, Wang HH. An economic framework of microbial trade. *PLoS ONE*. (2015) **10**:e0132907. doi: 10.1371/journal.pone.0132907
70. Varian HR. *Intermediate Microeconomics: A Modern Approach (Eighth Edition)*. 8th ed. W. W. Norton & Company (2009). Available online at: http://www.worldcat.org/isbn/0393934241
71. Werner GD, Strassmann JE, Ivens AB, Engelmoer DJ, Verbruggen E, Queller DC, et al. Evolution of microbial markets. *Proc Natl Acad Sci USA*. (2014) **111**:1237–44. doi: 10.1073/pnas.1315980111

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.

Copyright © 2019 Pusa, Wannagat and Sagot. 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) and the copyright owner(s) 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 Genetics

REVIEW
published: 11 April 2019
doi: 10.3389/fgene.2019.00294

Network Medicine in the Age of Biomedical Big Data

Abhijeet R. Sonawane[1,2], *Scott T. Weiss*[1,2], *Kimberly Glass*[1,2] *and Amitabh Sharma*[1,2,3]*

[1] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, United States, [2] Department of Medicine, Harvard Medical School, Boston, MA, United States, [3] Center for Interdisciplinary Cardiovascular Sciences, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA, United States

Network medicine is an emerging area of research dealing with molecular and genetic interactions, network biomarkers of disease, and therapeutic target discovery. Large-scale biomedical data generation offers a unique opportunity to assess the effect and impact of cellular heterogeneity and environmental perturbations on the observed phenotype. Marrying the two, network medicine with biomedical data provides a framework to build meaningful models and extract impactful results at a network level. In this review, we survey existing network types and biomedical data sources. More importantly, we delve into ways in which the network medicine approach, aided by phenotype-specific biomedical data, can be gainfully applied. We provide three paradigms, mainly dealing with three major biological network archetypes: protein-protein interaction, expression-based, and gene regulatory networks. For each of these paradigms, we discuss a broad overview of philosophies under which various network methods work. We also provide a few examples in each paradigm as a test case of its successful application. Finally, we delineate several opportunities and challenges in the field of network medicine. We hope this review provides a lexicon for researchers from biological sciences and network theory to come on the same page to work on research areas that require interdisciplinary expertise. Taken together, the understanding gained from combining biomedical data with networks can be useful for characterizing disease etiologies and identifying therapeutic targets, which, in turn, will lead to better preventive medicine with translational impact on personalized healthcare.

Keywords: network medicine, biological networks, biomedical big data, interactome, co-expression, gene regulations, phenotype-specificity, systems medicine

OPEN ACCESS

Edited by:
Marco Pellegrini, Italian National Research Council (CNR), Italy

Reviewed by:
Shailendra Kumar Gupta, University of Rostock, Germany
Adriano Velasque Werhli, Fundação Universidade Federal do Rio Grande, Brazil

***Correspondence:**
Amitabh Sharma
amitabh.sharma@channing.harvard.edu;
amitabhsharmaa@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 December 2018*
Accepted: *19 March 2019*
Published: *11 April 2019*

Citation:
Sonawane AR, Weiss ST, Glass K and Sharma A (2019) Network Medicine in the Age of Biomedical Big Data. Front. Genet. 10:294. doi: 10.3389/fgene.2019.00294

INTRODUCTION

Biological systems are comprised of various molecular entities such as genes, proteins and other biological molecules, as well as interactions between those components. Understanding a given phenotype, the functioning of a cell or tissue, etiology of disease, or cellular organization, requires accurate measurements of the abundance profiles of these molecular entities in the form of biomedical data. Analysis of the biomedical data allows us to explain important features of the interactions leading to a mechanistic understanding of the observed phenotype. The interplay between different components at different levels can be represented in the form of biological

Abbreviations: CNV, copy number variation; ENCODE, ENCyclopedia Of DNA elements; FANTOM5, Functional ANnoTation Of Mammalian Genome; GCNs, gene co-expression networks; GRNs, gene regulatory networks; GTEx, genotype-tissue expression; HCA, human cell atlas; HMP, human microbiome project; HPA, human protein atlas; modENCODE, model organism ENCyclopedia Of DNA Elements; NGS, next generation sequencing; PPIs, protein-protein interactions; SNP, single nucleotide polymorphism; TCGA, the cancer genome atlas; TOPMed, *trans*-omics for precision medicine.

networks, for example, protein-protein interactions (PPIs) (Uetz et al., 2000; Cusick et al., 2005) and gene regulatory networks (GRNs) (Davidson, 2006). Different biological networks capture the complex interactions between genes, proteins, RNA molecules, metabolites and genetic variants in the cells of organisms. These networks, also interchangeably known as graphs, are representations in which the complex system components are simplified as nodes that are connected by links (edges) (Vidal et al., 2011). Networks provide a conceptual and intuitive framework to model different components of multiple omics data from the genome, transcriptome, proteome, and metabolome (**Figure 1**; Liu and Lauffenburger, 2009).

The convenient representation of the biological components in graphs led to the field of network biology – a discipline that studies holistic relationships between various biological components by combining graph theory, systems biology, and statistical analyses (Lindfors, 2011; Walhout et al., 2012). Moreover, the quantitative tools of network biology offer the potential to understand cellular organization and capture the impact of perturbations on these complex intracellular networks (Wang et al., 2011). Network Medicine is an extension of network biology with a set of focused goals related to disease biology, including understanding disease etiology, identifying potential biomarkers, and designing therapeutic interventions, including drug targets, dosage, and synergism discovery (Loscalzo et al., 2017). Research in network medicine heavily depends on large datasets for building models, making predictions and assessing their validity. The promise of network medicine research is to develop a more global understanding of how perturbations propagate in the system by identifying the pathways, sub-types of disease states, and key components in the networks that can be targeted in clinical interventions. Moreover, networks are the centerpiece of the "new biology" in the biomedical data revolution and translation to personalized medicine (Schadt and Bjorkegren, 2012).

Advances in high-throughput biotechnologies have led to the generation of massive amounts of biomedical data that provides new research avenues. The rapid decline in costs due to technological advancements such as next generation sequencing (NGS) have provided the necessary impetus to generate multiple large-scale multi-omics biomedical data-sets that characterize various phenotypes. This includes exome and whole genome sequencing, transcriptomics, proteomics, lipidomics, microbiomics, etc. (Schadt and Bjorkegren, 2012). Constructing appropriate network models is a challenging problem that heavily depends on the study design, the phenotype under study, the molecular entities measured, and the type and size of the data. The field of network medicine is largely discovery — rather than hypothesis — driven, uncovering previously unknown relationships and leading to the identification of new biomarkers. The statistical rigor of network predictions comes from the study design and the size of the datasets. Large-scale consortium-based efforts looking at the various aspects of human biology have allowed the application of network-based methods to uncover new insights into the molecular mechanisms of the given phenotype, such as tissue specificity or disease context. In this review, we first examine various large-scale biomedical datasets and types of biological networks as summarized by **Figure 1**. We then provide three paradigms in which biological networks can be combined with big biomedical data to understand the given phenotype.

BIOMEDICAL DATA SOURCES

Recent technological advancements in sequencing technologies, resulting in a reduction in cost per base pair, have heralded an era of massive data generation for different types of molecular profiles across a broad range of phenotypes and diseases. After the completion of the human genome project (Collins et al., 2003), the HapMap project (The International HapMap Consortium, 2003) created an extensive catalog of common human genetic variants, the differences in DNA sequences, based on microarray data. These studies eventually progressed into the "1000 Genomes Project" (The 1000 Genomes Project Consortium, 2015), which leveraged NGS technologies. In cancer research, the cancer genome atlas (TCGA) (Cancer Genome Atlas Research Network, 2008) contains profiles of tumors and matched normal samples from more than 11000 subjects for 33 cancer types. The repertoire of TCGA data includes clinical information (demographic, treatment, and survival information), gene expression profiling, microRNA profiling, copy number variation (CNV) (genomic structural variations) identifications, single nucleotide polymorphism (SNP), DNA methylation (whole genome methylation calls for each CpG site), and exon sequencing (expression signal of particular composite exon of a gene). Together these data have helped in the identification of driver somatic mutations, the molecular basis of cancer progression, and potential therapeutic interventions for cancer subtypes. To understand the role of the epigenetic state in gene regulation and to characterize the functional elements of the transcriptional machinery, the ENCyclopedia Of DNA elements (ENCODE) consortium for humans (ENCODE Project Consortium, 2012), model organism ENCyclopedia Of DNA Elements (modENCODE) for model organisms (Yue et al., 2014), and ROADMAP Epigenomics project (Romanoski et al., 2015) were commissioned to improve the understanding of how epigenomics contributes to disease. The Riken-led Functional ANnoTation Of Mammalian Genome (FANTOM5) (Andersson et al., 2014) project provided cell-type-specific enhancer elements and identified pathobiological regulatory SNPs. To further understand transcriptional patterns in human tissues and their relationship with the genotype, genotype-tissue expression (GTEx) data was generated (GTEx Consortium, 2015; Mele et al., 2015). *Trans*-omics for precision medicine (TOPMed) (Prokopenko et al., 2018) is another set of multi-omics data on 100k individuals that also includes clinical data and is aimed at understanding the fundamental biological processes that underlie heart, lung, blood, and sleep disorders. The Precision Medicine Initiative or "All of Us" program[1] aims to acquire a broad range of data from about 1 million individuals.

[1] https://allofus.nih.gov/

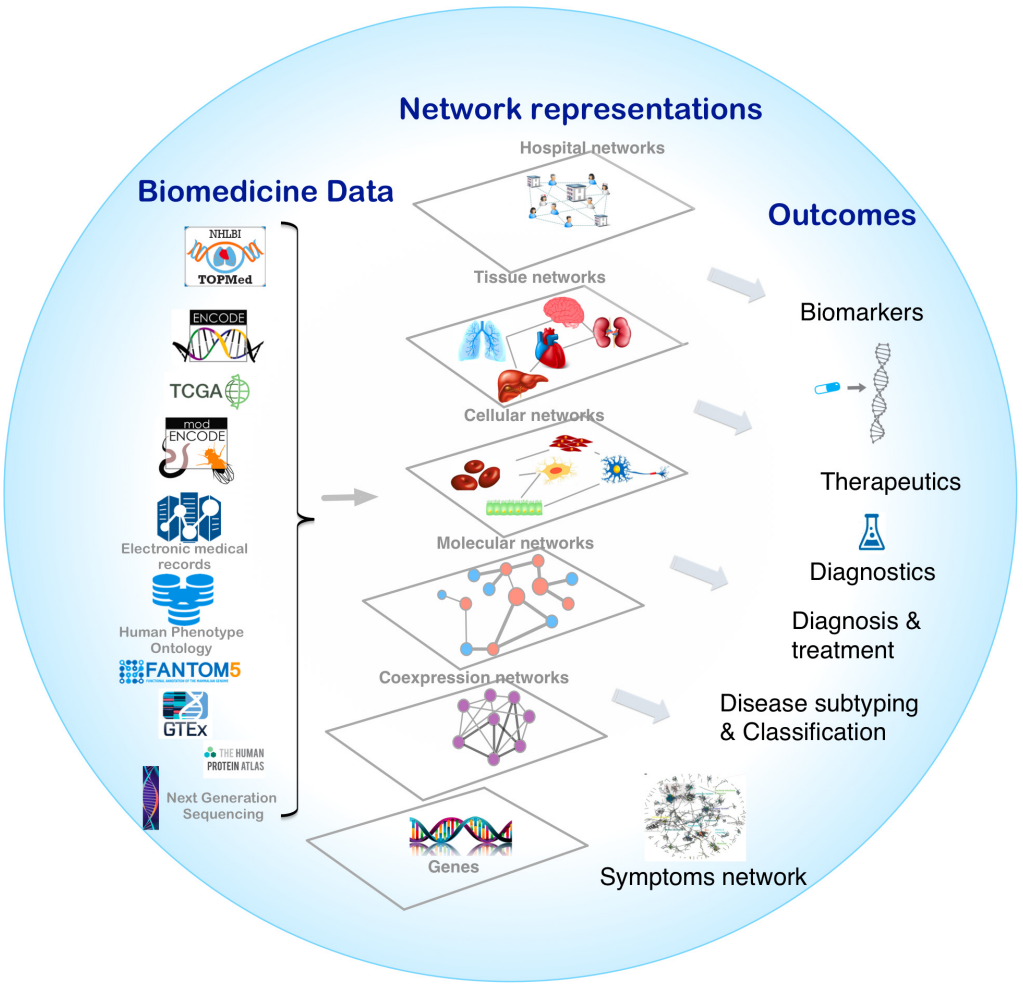


FIGURE 1 | Overview of network medicine approach depicting various biomedical data types discussed at length in the manuscript, along with network representations that simplify different components of multiple omics data from the genome, transcriptome, proteome, and metabolome as nodes that are connected by links (edges). Combining biomedical data with the appropriate network modeling approach allows derivation of disease associated information and outcomes like biomarkers, therapeutics targets, phenotype-specific genes and interactions, and disease subtypes.

Since 2003, the human protein atlas (HPA) (Uhlen et al., 2005; Uhlen et al., 2015), curated by Swedish consortium, has been releasing data on protein expression levels in cells, tissues, and various pathologies, including 17 cancer types. Similarly, the human cell atlas (HCA) (Rozenblatt-Rosen et al., 2017) aims to provide a reference map of single cell omics data in human cells and cell types. The UK-Biobank (Allen et al., 2014; Sudlow et al., 2015) is another commercial resource that has an array of health-related measurements on patients, including biomarkers, images, clinical information, and genetic data. The human microbiome project (HMP) (Turnbaugh et al., 2007) is a categorization of microbiota on different human body sites whose goal is to understand the role of the microbiome and the impact of its dysbiosis on human disease. Apart from these large international databases looking at one or more aspects of health or disease, many other resources from the concerted efforts over decades of data collection are also available. This includes the Nurses' Health Study (Belanger et al., 1978; Colditz et al., 2016), Health Professionals Follow-up Study (Grobbee et al., 1990), Framingham Heart Study (Dawber et al., 1951; Mahmood et al., 2014), and COPDGene (Pillai et al., 2009). This wealth of biomedical data not only allows for a deeper probing of the underlying biological systems, but also inspires the development of novel methods that can maximize the information that can be extracted from these data. The tools developed within the field of network medicine are highly versatile, enabling their customized application depending on the given biological or disease context.

Collecting large-scale multi-time point data across multiple omics in different disease conditions is expensive and often not feasible, especially for human subjects. However, small-scale longitudinal data for a single omic, such as gene expression, is available in biomedical databases (Jung et al., 2015; Bouquet et al., 2016). High resolution mass spectrometry has also allowed for the collection of longitudinal proteome data, for example to test the effect of drugs (Fournier et al., 2010) or oxidative stress (Vogel et al., 2011) in yeast. A longitudinal multi-omic

dataset containing both human transcriptomic and proteomic information has been analyzed to study changes in molecular profiles (Chen et al., 2012). Multi-omic datasets such as this one allows us to probe the relationship between biological molecules based on the central dogma of biology, such as the connection between transcript abundance and protein levels (Marguerat et al., 2012; Liu et al., 2016). Longitudinal data is also amenable to temporal or dynamical network analysis, wherein one can evaluate the statistical dependence of the state of a network on the gene expression patterns from previous time steps (Kim and Kim, 2018; Dondelinger and Mukherjee, 2019). Kim et al. provide a summary of several methods to infer temporal regulatory relationships (Kim et al., 2014).

In the next section, we will review some of the main types of biological networks constructed using high throughput molecular profiling, literature mining, or manual curation of scientific literature.

PRIMER ON BIOLOGICAL NETWORKS

Each network-based study has to primarily identify two things: what are the critical entities in the system under investigation (nodes), and what is the nature of the interactions between these entities (edges) (de Silva and Stumpf, 2005). This information often comes from multiple different data sources, dealing with the various facets of the biological system. For example, PPIs, also defined as the interactome, is a network of proteins and the physical interactions between them (Cusick et al., 2005). These interactions can be obtained from yeast-2-hybrid assays (Li et al., 2004; Vidal and Fields, 2014), co-immunoprecipitation (Lin and Lai, 2017), literature text-mining (Papanikolaou et al., 2015), 3D structure (Lu et al., 2013), co-expression of genes (Bhardwaj and Lu, 2005), sequence homology (Shen et al., 2007), and other sources. Each of these data sources have both merits and demerits (Cusick et al., 2005). These networks inform us about the overall topological properties of protein interactions as well as the positions of specific proteins within this network. However, extracting phenotype specific (i.e., cell, tissue or disease-specific) information based on the PPI remains an open challenge and requires the development of novel ways of integrating biomedical data with these networks.

Gene co-expression and regulatory networks often make direct use of phenotype-specific gene expression data in the network construction, with additional analysis required to extract meaningful biological information for the underlying phenotype. The availability of transcriptomic data for a wide range of phenotypes presents an opportunity to probe the patterns of molecular co-abundance, albeit with limitations concerning the interpretation of the biology. Gene co-expression networks (GCNs) can be constructed in many ways, including information theoretic, regression-based, and Bayesian approaches (Butte and Kohane, 1999). Several common methods for constructing GCNs include Weighted Gene Co-expression Network Analysis (WGCNA; Langfelder and Horvath, 2008), Context Likelihood of Relatedness (CLR; Faith et al., 2007), Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe; Margolin et al., 2006), Partial Correlation and Information Theory (PCIT; Reverter and Chan, 2008), Gene Network Inference with Ensemble of Trees (GENIE3; Huynh-Thu et al., 2010), Supervised Inference of Regulatory Networks (SIRENE; Mordelet and Vert, 2008), and Gene CO-expression Network method (GeCON; Roy et al., 2014). GRNs are a related type of network that attempts to look beyond the co-abundance of gene expression and instead identify the influencing patterns of transcription factor genes over others in a mechanistic fashion (Marbach et al., 2012). Since transcriptional regulation depends on *cis* and *trans*-regulatory elements as well as transcription factor binding, GRNs often incorporate this information during model construction. Many methods with a modified definition of correlations have been proposed to infer GRNs. However, identifying the putative *cis*-regulatory sequences, such as those found in the promoter regions of genes, that are relevant for a specific biological context is important to enable the understanding of disease, tissue, or cell-specific regulatory perturbations. The location of TF binding to the DNA can be assayed using yeast-1-hybrid (Deplancke et al., 2004), ChIP-Seq (Jaini et al., 2014), or inferred by other means (Mundade et al., 2014). However, the cost and other limitations involved in generating these data in a context-specific manner have meant that incorporating this information when constructing putative regulatory networks remains a challenge.

Other types of biological networks include metabolic networks, which represent a collection of biochemical interactions between metabolites and enzymes (Terzer et al., 2009). Ecological networks, which represent biotic interactions, can also be applied to microbiome data, the collection of microbes' genes, to construct microbiome networks (Coyte et al., 2015; Layeghifard et al., 2017; Bauer and Thiele, 2018; Rottjers and Faust, 2018). Together, genotype and transcriptomic data can be used to map genetic variants to genes and then summarized in an expression Quantitative Trait Loci (eQTL) network (Platig et al., 2016; Fagny et al., 2017). A network of immune cell communication has been constructed using high-resolution mass spectrometry-based proteomics data and was shown to exhibit social network-like properties. Disease networks, also known as the diseasome, have been proposed; these networks connect diseases and disorders with disease genes based on Online Mendelian Inheritance in Man (OMIM) associations (Boyadjiev and Jabs, 2000; Hamosh et al., 2002; Goh et al., 2007; Wysocki and Ritter, 2011; Zhang et al., 2011). Similarly, networks connecting symptoms with diseases have helped to shed light on the shared genetic associations between diseases (Zhou et al., 2014). Efforts to identify specific disease-causing genes, using genomic intervals obtained from linkage mappings or Genome-Wide Association Studies (GWAS), have been undertaken using hybrid heterogeneous networks. These hybrid networks often include a combination of disease-gene networks, generic or tissue-specific molecular networks such as PPIs or GCNs, and prior knowledge of disease similarities (Navlakha and Kingsford, 2010; Moreau and Tranchevent, 2012; Ni et al., 2016). Various network-based tools have been implemented in the gene prioritization problem (Wu et al., 2008; Li and Patra, 2010; Tian et al., 2017). All these aforementioned types of network

biology approaches are particularly useful in understanding complex diseases, which result from multiple genetic factors and environmental influences (Moreau and Tranchevent, 2012).

Analysis of biological networks also necessitates understanding their structural or topological properties. This includes the identification of important modulators, driver nodes, local network structures, and recurrent subgraphs in the network. Local connectivity properties such as degree and other centrality metrics can help to identify key molecular entities that dominate various network neighborhoods, such as hubs, bottlenecks, or core nodes. At the global level, properties like average path length, degree distribution, diameter, clustering coefficients, and controllability (Liu et al., 2011) help with the characterization and comparison of network topologies. Mesoscale measures such subgraphs or network motifs – recurrent patterns connecting a fixed number of nodes (typically 3 or 4) – are considered fundamental components of biological networks (Milo et al., 2002). An extension of network motifs to include more nodes, or graphlets, has been used to analyze the interactome (Przulj et al., 2004; Davis et al., 2015; Malod-Dognin et al., 2017). Identifying the connectivity patterns enriched in a network (i.e., over-represented with respect to a null model) can help to compare, characterize, and discriminate between networks (Shen-Orr et al., 2002; Alon, 2007; Przulj, 2007). These patterns are also commonly associated with control substructures that dominate information flow in the networks, especially in transcriptional regulatory, neuronal, and social networks.

INTEGRATING BIOMEDICAL DATA WITH NETWORKS: CHALLENGES AND WAYS

The ultimate aim of inferring biological networks using biomedical data is to provide lab-testable hypotheses by identifying biomolecular entities that play a crucial role in the observed phenotype (**Figure 1**). Detecting changes in abundance levels of these biomolecules and their interaction landscape in the context of a tissue, cell, or disease-specific environment requires both relevant data and the application of appropriate network analysis. Each biological network analysis has strengths and limitations based on how it incorporates phenotype specific data, and the research question being addressed (Altaf-Ul-Amin et al., 2014; Kanaya et al., 2014). In some cases, it is possible to identify a baseline network from general physical interactions between proteins, after which disease or phenotype-specific information from specific experiments can be overlaid to generate a more context-specific network.

Protein-protein interaction networks provide a fabric of potential interactions between proteins, but phenotype-specific interactions can only be added as an extra layer from separate biomedical data. The hypothesis behind analyzing such networks, combination of baseline PPI with disease information added as next step, is that the defects or mutations in only a few genes or proteins may propagate to other components in the network, and that this collection of affected genes constitute a critical module in the network (Schadt and Bjorkegren, 2012). Previous work along these lines has shown that these modules are not only structurally related but are also functionally relevant to the observed phenotype. This central tenet of network medicine from the interactome has been successfully tested for many diseases and other phenotypes (Lim et al., 2006; Goh et al., 2007; Taylor et al., 2009; Sharma et al., 2013, 2015, 2018; Menche et al., 2015; Sahni et al., 2015; Huttlin et al., 2017; Huang J.K. et al., 2018; Wang et al., 2018; Willsey et al., 2018) and has also led to novel drug-target discoveries (Yildirim et al., 2007; Guney et al., 2016; Luo et al., 2017) along with novel interactions between genes. Despite recent advances, the PPI is incomplete and inferring disease-specific interactions requires innovative strategies in order to overcome this deficiency.

Gene co-expression networks are by definition context-specific, as they are constructed by calculating correlations in a given gene expression data set. In contrast, GRNs often are built starting from a baseline network composed of all potential interactions between transcription factors and genes. This baseline network can be derived from genetic sequence information and DNA-binding domain sequences within regulatory proteins, such that an interaction is inferred if a given gene's promoter contains the binding motif of a particular TF. Disease or tissue-specific information then has to be integrated with this baseline prior network to obtain meaningful information about perturbations caused due to the disease.

In this review, we explore the PPI, GCNs, and GRNs, and also provide exemplar methods for each. Based on these three types of networks, we describe three complimentary philosophies and *modus-operandi* to embed phenotypic specific molecular information from biomedical data into a network framework, as shown in **Figure 2**. We present these paradigms to demonstrate that applying network phenomenology to big biomedical data requires a nuanced, condition-specific approach. In the following sections, we will focus on each paradigm separately, providing their examples, the questions they intend to answer, and the diagnostics of the outcomes. We mainly focus on reviewing methods to integrate multi-omic data to extract phenotype specific information, specifically disease and tissue specificity in the PPI, GCNs, and GRNs.

PARADIGM I: Network-Based Approach to Human Disease Using the Interactome

The high-throughput mapping of the interactome has provided a molecular interaction map of the genes encoding proteins that might drive an underlying pathophenotype (Kamburov et al., 2009; Barabasi et al., 2011; Zhang et al., 2013; Rolland et al., 2014; Hein et al., 2015; Huttlin et al., 2015). Understanding disease associated biomedical data in the context of network principles supports the discovery of more accurate biomarkers, localization of the disease perturbation in the network, personalized networks, better disease sub-type classifications, better targets for drug development, and better drug repurposing. Using this paradigm, one can extract disease-specific signals in a variety of ways. One may consider topological properties of the nodes and assess the functional role of their *hubness*, i.e., a node property of having a higher number of connections. Alternatively, one can

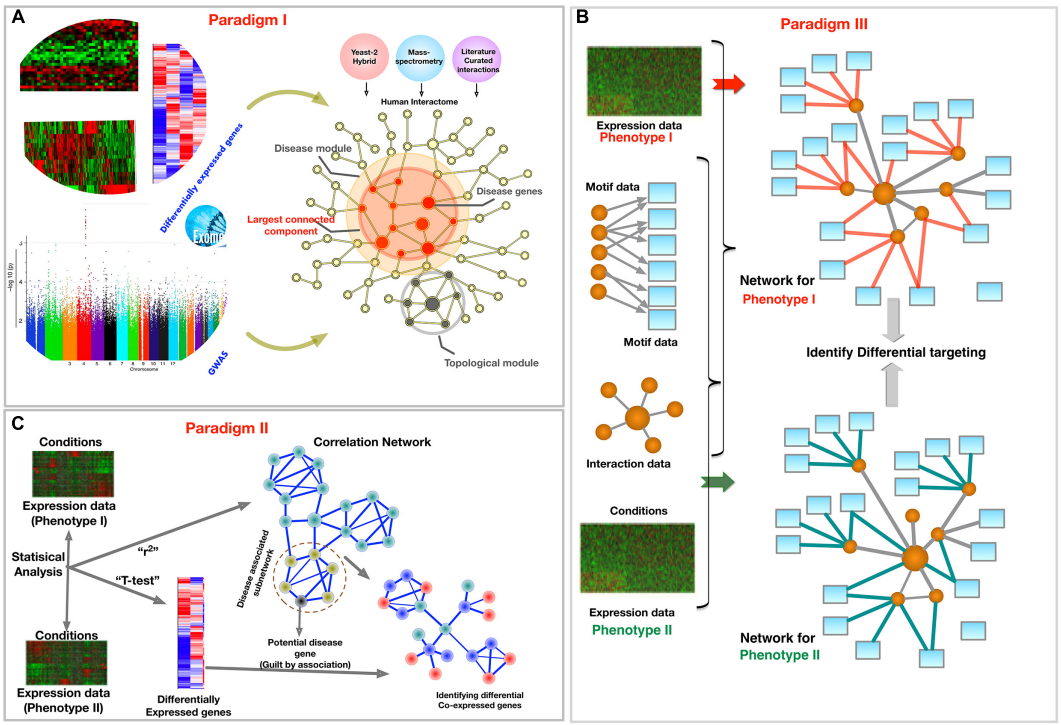


FIGURE 2 | Schematic of three paradigms for combining biological networks with phenotype-specific biomedical data, such as a set of disease genes and transcriptomic profiles for case and control groups. **(A)** Identification of disease associated network components within the interactome, **(B)** Co-expression based network modeling to identify disease biomarkers, **(C)** Constructing phenotype-specific GRNs to identify perturbations and condition-specific regulatory changes.

also identify new disease genes in the network by using "guilt-by-association" (Aravind, 2000; Quackenbush, 2003; Stuart et al., 2003; Lage et al., 2007; Sharma et al., 2010; Lee et al., 2011; Sharma et al., 2013; Huang J.K. et al., 2018) — a property ascribed not based on direct evidence but association with other disease genes, albeit with care (Gillis and Pavlidis, 2012). In addition to prioritizing candidate disease genes, molecular interaction networks can assist in identifying the sub-networks that are mechanistically linked to disease phenotypes (Menche et al., 2015; Sharma et al., 2015; Emamjomeh et al., 2017; van Dam et al., 2018). The proteins in these connected subnetworks may have clinical importance by being therapeutic targets and biomarkers (Sharma et al., 2015). Network tools can also provide a framework for disease classification (Halu et al., 2017; Zhou et al., 2018).

Assessing disease genes from other, non-disease genes by their topological properties on the interactome have provided new insight into disease pathobiology. It was found that disease genes tend to have non-hub properties (Goh et al., 2007). Later, it was reported that genes from OMIM and those associated with cancer are more central in a literature-curated interactome (Jonsson and Bates, 2006; Xu and Li, 2006; Ideker and Sharan, 2008). Further, several studies demonstrated that disease genes, in general, mostly have a high-degree and a low clustering coefficient (number of mutual connections with the neighboring nodes) (Feldman et al., 2008; Cai et al., 2010). Moreover, recently it was reported that disease genes have a higher degree, but it was discovered that the cancer-related genes are the primary drivers of this trend (Wachi et al., 2005; Jonsson and Bates, 2006). Genes associated with either Mendelian or complex diseases also have higher degree and lower clustering coefficients compared to non-disease genes (Cai et al., 2010; Pinero et al., 2016). The topological properties of disease-associated genes vary significantly from disease to disease. The factors that influence these discrepancies include the incompleteness of the current interactome, bias toward well-studied genes, and incomplete knowledge about the number genes associated with various diseases (Menche et al., 2015). It is anticipated that the alliance of different technologies like yeast-2-hybrid, affinity purification mass-spectrometry (AP-MS), and cross-linking AP-MS (Schweppe et al., 2018) will provide access to larger data that will be helpful in providing knowledge about the missing interactions. On the disease-gene discovery side, projects like the UK biobank prospective cohort study, which includes in-depth genetic and phenotypic data, will enhance knowledge regarding the missing disease genes (Bycroft et al., 2018).

An important area in which the interactome has helped in understanding complex diseases is the prediction of disease associated genes. The goal is to identify novel genes and proteins, which are involved in the regulation of tissues, or dysregulated in the case of disease, through the association with observed disease candidate genes using the biological hierarchy of molecular interactions. **Figure 2A** depicts this paradigm where the PPI network serves as map of potential biological interactions between various proteins over which disease associated genes

are mapped to uncover relevant biology. The central philosophy in most methods under this paradigm is that the neighbors of the disease associated components or network modules, such as a set of differentially expressed genes (Chuang et al., 2007) or genes with disease-associated SNPs (Oti et al., 2006; Lage et al., 2007; Feldman et al., 2008; Barrenas et al., 2012), could potentially be associated with similar diseases (Goh et al., 2007), and are closer to each other as compared to the other nodes in the network. The definition of this closeness, or vicinity of nodes, just like the definition of modules and clusters, varies with different research strategies. Some methods assume topological closeness in terms of the number of shortest paths connecting given nodes, while others take the similarity of biological function into account. Guilt-by-association methods focus on identifying new disease genes by optimizing based on both the local and global properties of the network and by considering the role of other disease genes and their neighborhood. Network-based strategies to find disease genes and their associated mechanisms can be divided in two types: exploratory and analytic methods (Carter et al., 2013). In exploratory methods one can analyze the biological trends due to perturbations. For example, Chu et al. (2012) expanded on known angiogenesis pathways to construct a PPI network for angiogenesis. In contrast, analytic methods aim to identify specific genes and pathways associated with a disease. For example, Gilman and group developed a method for network-based analysis of genetic associations to identify a biological network of genes affected by rare *de novo* CNVs in autism (Gilman et al., 2011). Recently, Huang J.K. et al. (2018) systematically evaluated 21 protein-interaction networks for the ability to recover disease genes sets. After correcting for size, they found that the Database for Interacting Proteins (DIP) network (Xenarios et al., 2000) had the highest efficiency in recovering disease genes (Huang J.K. et al., 2018).

In contrast to predicting the disease candidate proteins, finding the associated disease-related network components, or sub-networks, provides a more substantial network space to discover the pathways and mechanisms that influence disease. Goh et al. (2007) proposed a correlation between the location of disease-associated genes and the topology of the molecular interaction network. The tendency of disease-associated genes to interact more often with others compared to random genes in the interactome led to the establishment of the 'local impact' hypothesis (Barabasi et al., 2011). According to this hypothesis, molecular entities involved in similar diseases have an increased tendency to interact with each other and to localize in a specific neighborhood of the interactome (Barabasi et al., 2011). The search for these modules involves exploring the structural and topological properties of the PPI network. Community detection algorithms (Spirin and Mirny, 2003), clique percolation (Sun et al., 2011), and genetic algorithms (Liu et al., 2018) have been applied to uncover disease modules using network properties (Vlaic et al., 2018). Module prediction and identifying non-overlapping clusters with the PPI remains challenging since the PPI network has a short diameter, i.e., most nodes are close to all other nodes in terms of network distance. Novel distance metrics and community detection algorithms have been proposed to overcome this problem (Hall-Swan et al., 2018). The recently proposed DIseAse MOdule Detection (DIAMOnD) algorithm (Ghiassian et al., 2015) associates the functional modules of known disease-associated proteins (seed proteins) and identifies the close neighbors of these genes (candidate disease-associated proteins) using topological properties of the interactome. The method suggests that the *connectivity significance* among the disease-associated proteins is the best predictive quantity to find the disease related components in the interactome. The underlying hypothesis is that close neighbors of known disease proteins may be involved in the disease. The working principle of DIAMOnD is as follows: first, a pool of disease genes encoding proteins is identified for a disease of interest from biological experiments, GWAS, linkage analysis, or other disease associated data sources (Pinero et al., 2017). Next, these disease proteins (seeds) are mapped onto the interactome. Further, neighbor proteins are added iteratively to the set of seed proteins based on the condition that each neighbor added is most significantly connected to the seed proteins. A hypergeometric test assigns a p-value to the proteins that share more connections with seed proteins than expected by chance. Finally, the seed proteins plus the added neighbor proteins are part of network components that represent a disease module, or a subnetwork of proteins in the interactome, the members of which are more functionally and topologically related to each other than to other portions of the network. These subnetworks are designated as disease-specific modules based on the source of initial seed proteins. Disease module identification has also led to endophenotypes, intermediate pathophenotypes, and network modules describing their common and distinctive molecular mediators (Lage et al., 2008; Ghiassian et al., 2016).

As mentioned previously, significant progress has been made in mapping the interactome by high-throughput approaches like yeast-2-hybrid (Rual et al., 2005; Venkatesan et al., 2009; Dreze et al., 2010; Rolland et al., 2014), AP/MS (Hein et al., 2015; Huttlin et al., 2015, 2017) and various literature-curated data sources, such as ConsensusPathDB, STRING, and PCNet, which collate the known and predicted interactions between proteins (Klingstrom and Plewczynski, 2011). Despite these efforts, the current interactome mapping is 80% incomplete (Hart et al., 2006; Venkatesan et al., 2009; Mosca et al., 2013; Menche et al., 2015) and is affected by many experimental and literature biases. Given the incompleteness of the interactome, it is possible that the disease modules are also far from complete. An attempt to overcome this limitation was made using a network-based closeness approach that compares the weighted distance between different disease and seed-gene neighborhoods to random expectation on the network. In the context of Chronic Obstructive Pulmonary Disease (COPD), 140 potential candidate genes (Sharma et al., 2018) were identified. Another shortcoming of disease module detection related to the lack of context-dependence and tissue-specificity within the PPI was studied by Kitsak et al. (2016). They found that the genes expressed in a particular tissue tend to form localized connected subnetworks, which overlap between similar tissues and are situated in the different neighborhoods for pathologically distinct pairs of tissues. The perturbations in tissue-dependent subnetworks may help us understand disease manifestations

or pathophenotypes. Integrating multi-omics data, including epigenomics, proteomics, and metabolomics, with PPI analysis remains challenging, but is critical for identifying disease or tissue-specific modules in the interactome.

PARADIGM II: Identifying Important Genes Using Patterns of Co-abundance of Biomolecules

Measuring transcript abundance or gene expression patterns for given phenotypes (case-control) across multiple samples is one of the main research strategies used to probe the system as it is connected to the central dogma of molecular biology. Performing differential gene expression analysis often identifies important genes affected by the disease. However, it does not provide information regarding how these genes are influenced by or influence other genes. It has been observed that genes with similar expression patterns might be part of complexes, influence each other, or be part of the same pathways or mechanisms (Serin et al., 2016). This inspired the construction of GCNs where the patterns of transcript abundance are studied in the context of the disease. The central philosophy of this paradigm is to combine important seed genes with an organic network of co-expression patterns derived from the gene expression data from the same system.

There are many ways to compute co-expression or co-abundance patterns, including using Pearson correlations (Stuart et al., 2003), Spearman rank correlations (Song et al., 2012; Liesecke et al., 2018), mutual information (Butte and Kohane, 1999; Margolin et al., 2006; Meyer et al., 2007), Gaussian graphical models (Toh and Horimoto, 2002), regression-based methods (Yeung et al., 2002; van Someren et al., 2006; Pirgazi and Khanteymoori, 2018), Bayesian approaches (Friedman et al., 2000; Perrin et al., 2003; Li et al., 2007; Xing et al., 2017), random matrix theory (Luo et al., 2007; Jalan et al., 2010; Jalan et al., 2012), and partial correlations (Reverter and Chan, 2008). GCNs identify the functionally coordinated participation of genes in response to an external stimulus or condition. GCNs can be signed or unsigned, weighted or unweighted, and may either be constructed using microarray or RNA-Seq data. Care must be exercised when using thresholding methods to obtain unweighted co-expression networks as these are subjective and can change the network structure and topology (Elo et al., 2007); methods based on the clustering coefficient (Boyadjiev and Jabs, 2000), random matrix theory (Luo et al., 2007), or soft thresholding, which raises the weights by a certain power to penalize weaker edges (Langfelder and Horvath, 2008), have been used to address this limitation. Along with total gene expression levels, isoform abundance and alternative splicing can also be used in constructing GCNs (Saha et al., 2017).

Gene co-expression networks are also used to identify co-expression modules. Clusters, modules, or subgraphs of genes that have similar functions are often highly interconnected in GCNs. These clusters can be identified using network topology-based methods like community detection (Girvan and Newman, 2002), modularity maximization (Newman, 2004), K-means clustering (Stuart et al., 2003), or variants of hierarchical clustering methods (Langfelder and Horvath, 2008; Serin et al., 2016). The genes in the most significant modules are then assessed for their biological importance using functional enrichment methods. The genes in the clusters are also often tested for their enrichment with differentially expressed genes from transcriptomic analysis, as illustrated in **Figure 2B**. Based on these results, other non-differentially expressed genes in the enriched clusters can be implicated in the disease using 'guilt-by-association' approaches. The newly implicated genes may have clinical importance as potential therapeutic targets and biomarkers.

Despite the aphorism "correlation is not causation", partial yet informative insights can be gleaned from co-expression networks, such as an underlying regulatory framework mediating the co-expression patterns. New methods based on partial-correlations, Bayesian, and graphical Gaussian models (Werhli et al., 2006) take into account local connectivity when estimating edge strengths and a few methods work by combining prior-knowledge of expression patterns of TFs with co-expression information (Huynh-Thu et al., 2010; Rotival and Petretto, 2014). Gene-gene interaction network methods like ARACNe (Margolin et al., 2006) and CLR (Faith et al., 2007) attempt to better capture these regulatory associations by accounting for connections within a shared neighborhood of genes in order to infer the strength of a link between two genes. Applying these approaches in complex conditions, like a gene being regulated by many regulators, becomes more challenging. Inferring the direct regulatory influence of transcription factors on target genes is central to interpreting the regulatory networks. Concerted efforts to support network-inference, such as the DREAM5 benchmark challenge (Marbach et al., 2012), have summarized different strategies that can be employed to infer regulatory networks. The accuracy of reconstruction approaches is often tested by comparing the predicted networks with high-confidence transcription factor binding data (He and Tan, 2016). However, integrating multi-omic data into these models to understand the pathobiology of disease states is an open challenge. Methods like CellNet (Cahan et al., 2014), an extension of CLR, and MOGRIFY (Rackham et al., 2016) take into account differentially expressed genes within the co-expression network framework in order to predict cellular reprogramming by transcription factors. Thus, co-expression methods have also been used to infer regulatory networks and to delineate the influence of regulatory genes, such as transcription factors, on their targets. However, obtaining condition-specific GRNs requires information regarding transcription factor binding activity in the given context. We will review some of the methods that utilize TF binding information in the next section.

To summarize, inferring disease-specific information from GCN is possible from co-expressed or co-regulated clusters, differentially expressed and co-expressed genes, as well as the topological and functional properties of these. Biomedical big data measuring the transcriptome is highly leveraged by GCNs. For example, human tissue-specific GCNs have been constructed and analyzed (Pierson et al., 2015) using consortium data such as GTEx (Mele et al., 2015). These analyses revealed that genes with tissue-specific function are not hubs but connect to tissue-specific

transcription factor hubs. Explorations using relative isoform ratios (RNA transcripts from the same genes with different exons removed) and splicing data revealed distinct co-expression relationships unique to the tissues (Saha et al., 2017). Tissue specificity of GCNs have also been assessed in rats (Xiao et al., 2014), humans (Prieto et al., 2008; Xiao et al., 2014; Kogelman et al., 2016; Ni et al., 2016; Farahbod and Pavlidis, 2018), bats (Rodenas-Cuadrado et al., 2015), and plants (Aravind, 2000). Similarly, TCGA data has been analyzed using WGCNA in order to study the system-level properties of prognostic genes (Yang et al., 2014). Similar to gene co-expression, protein co-abundance networks can also be used to pinpoint influential proteins as potential regulators of the observed phenotype, and have been used to study inflammation (Halu et al., 2018), HCV infections (McDermott et al., 2012), and cancer, including breast cancer (Ryan et al., 2017) and glioblastoma (Kanonidis et al., 2016).

PARADIGM III: Inferring Phenotype Specific Gene Regulatory Networks

In the previous sections, we studied various ways to construct networks and integrate molecular data to extract phenotype-specific biology in the form of gene prioritization, disease modules, or therapeutic targets. Those included immutable PPIs allowing disease-specific information to be embedded onto them and organic ways to model disease-specific information using co-expression networks. Here, separate networks are built for each phenotype which may be case-control, disease-specific, tissue or cell-specific, sex-specific, or for different disease subtypes. The network comparison model stems from the axiom of "differential networking" over "differential expression." Many examples of differential networking can be found, including the INtegrated DiffErential Expression and Differential network analysis (INDEED) (Zuo et al., 2016) and DICER (Amar et al., 2013) algorithms. In this paradigm, we aim to discuss ways of leveraging phenotype-specific biomedical information to construct condition-specific GRNs. In principle, GCNs can also be phenotype-specific and can be used to infer condition-specific signals, but they lack the underlying set of canonical interactions unlike GRNs which include protein-DNA interaction in the form of TF binding information.

Instead of combining data from cases and controls to obtain key molecular elements, such as differentially expressed genes or genes annotated to GWAS SNPs, in this paradigm the data is used to construct separate networks for each of the conditions. This construction of phenotype specific networks helps to mitigate systematic experimental biases and errors in both conditions (de la Fuente, 2010; Ideker and Krogan, 2012). It allows the comparison of networks to help uncover the specific rewiring of pathways, such as those induced by disease, pharmacological treatment (Bandyopadhyay et al., 2010), or environmental stimuli. GCNs can also be constructed in a phenotype-specific manner, as seen in the previous section. In **Figure 2C**, we depict an approach where phenotype-specific networks are constructed to uncover differentially targeted interactions. In this section, we focus on transcriptional regulatory networks that depend not only on co-expression, but also on modeling the binding propensities of TFs. These networks may also incorporate other multi-omic data to obtain condition-specific regulatory models.

The primary benefit of comparing phenotype-specific networks, particularly in GRNs, is to better delineate the role of genes in each condition. The "rewiring" of the TFs targeting each of the genes can be tracked and the perturbations leading to these changes can convey information regarding the mechanistic underpinnings of the observed phenotype. An apt extension of "differential networking" to the transcriptional regulatory network framework is "differential targeting," which captures the highly dynamic nature of gene regulation. Changes in network topology, driven by underlying condition-specific data, can yield valuable insights and help to identify driver nodes and network biomarkers, such as a set of strengthened or weakened interactions between TF and target genes in the context of disease.

We review the Passing Attributes between Networks for Data Assimilation (PANDA) algorithm (Glass et al., 2013) as an exemplary method for constructing condition-specific regulatory networks, allowing for robust differential targeting analysis. PANDA is initiated by constructing a prior regulatory network consisting of potential routes for communication by mapping transcription factor motifs to a reference genome and assigning them to genes if they are in the regulatory region of the genes. PANDA then integrates other sources of information to iteratively optimize the flow of information through the network, modifying the prior to obtain a condition-specific regulatory network. The phenotype-specific regulatory networks are then compared to identify the structures most affected by this "rewiring" and their biological significance. PANDA models the interactions between transcription factors based on the following principles. Firstly, if two transcription factors have a similar targeting profile, i.e., target similar genes or have binding motifs in the promoters of the same genes, they are more likely to physically interact or be members of the same TF complex (Hemberg and Kreiman, 2011; Guo et al., 2016). Cooperative binding of TFs is found to be evolutionarily constrained and conserved (Goke et al., 2011; He et al., 2011), and impacts crucial eukaryotic functions (Hochedlinger and Plath, 2009; Wilson et al., 2010; He et al., 2011; Will and Helms, 2014). Likewise, if two genes are targeted by the same set of TFs, these genes are likely to share similar expression patterns (Yu et al., 2003; Kim et al., 2006; Marco et al., 2009), or be part of the same functional module (Goh et al., 2007; Feldman et al., 2008). For this purpose, PANDA incorporates PPI networks to determine the "responsibility" of TFs co-binding based on shared targets. It also uses GCNs to determine the "availability" of genes to be simultaneously co-regulated, as evidenced by common co-expression. A vital component in PANDA is a "prior" network composed of all potential regulatory routes based on the existence of binding sites for TFs in the regulatory regions of genes. All three ingredients (PPI, GCN, and a network prior) are then assimilated to uncover consistent patterns among these networks using a message-passing framework similar to affinity-propagation (Frey and Dueck, 2007). The outcome is a network elucidating the edges that form self-consistent modules, identifying relevant biological processes.

The phenotype-specific applications of PANDA are broad and include the comparison of disease and control networks in both complex diseases and cancers. For example, PANDA has been used to identify potential drug targets in ovarian cancer subtypes (Glass et al., 2015). Comparing PANDA networks between poor and good responders to asthma therapies identified potential transcriptional mediators of corticosteroid response in asthma (Qiu et al., 2018). The role of serotonin (5HT) dysregulation in mitral valve disease was explored using PANDA to find upregulation in 5HTR2B expression and an increase 5HT receptor signaling (Driesbaugh et al., 2018). The effect of weight-loss on decreased risk of colorectal cancer was evaluated by applying PANDA to gene expression data on rectal mucosa biopsies (Vargas et al., 2016). In cancer research, PANDA network analysis in triple-negative breast cancer (TNBC) identified new core modules of functionally essential TFs and genes in cancer cells (Min et al., 2017). PANDA has also been used to investigate non-epithelial cancers like glioma to identify prognostic biomarkers mainly concerning mesenchymal signatures (Celiku et al., 2017). Sexual dimorphism, where the phenotypes are males and females, is another area where PANDA has been applied extensively, from sex-related targeting differences in COPD (Glass et al., 2014), colorectal cancer (Lopes-Ramos et al., 2018), and understanding crucial sex-related differences in various tissues in the human body (Chen et al., 2016). Differences between cell-lines and their host tissues have also been investigated using PANDA (Lopes-Ramos et al., 2017).

The issue of tissue-specificity can also be addressed by the paradigm of condition-specific networks, where the phenotype is the tissue or cell type. Various methods use gene expression data with regression trees (Huynh-Thu et al., 2010) or consider the context of pathways (Jambusaria et al., 2018). Enhancer and promoter data (Marbach et al., 2016) have been used to construct tissue-specific networks in humans and plants (Huang J. et al., 2018). Using GTEx transcriptome data, PANDA has been used to construct GRNs for 38 distinct human tissues (Sonawane et al., 2017). This analysis assessed the inter-relationship between tissue-specific genes and TFs based on expression data and tissue-specific interactions and the topological positions of functionally important genes in respective tissues. This study also used network centrality measures like betweenness and degree to assess the topological properties of the nodes to identify rewiring around these genes in various tissues. Another significant contribution of this work is the elucidation of the tissue-specific regulatory roles of transcription factors, which were found to be independent of their expression levels. Instead, transcription factors appeared to mediate critical tissue-specific processes through subtle shifts in the GRNs, providing functional redundancy and, as a consequence, phenotypic stability of tissues.

CONCLUSION AND FUTURE DIRECTIONS

Above we reviewed a limited set of network medicine philosophies that seek to integrate biomedical big data to uncover meaningful biology. Network medicine approaches provide customized and optimized ways to leverage biomedical data. The choice of the appropriate network method is largely dictated by the underlying biological inquiry, hypotheses, study design, and available data. Although this review is not meant to be exhaustive, our intent was to give a essence of how biomedical data requires a nuanced approach when selecting network analyses and provide a resource for both network scientists and biologists to better understand the lexicon of network modeling of biomedical data.

We believe that network medicine approaches will be vital in the future with the increasing emergence of diverse technologies, multi-omic data types, deeper levels of inquiry from tissues to cellular levels, platforms that include large amounts of publicly available biomedical data, and efforts in precision medicine, which aim to find the right drugs for the right patients at the right time. There is a growing realization that genomics is only a part of the story when it comes to cancer and other complex diseases. The field is working to augment genetic information (mutations, deletions, and other somatic genetic alterations) with other omics data, such as epigenomics (methylation, non-coding RNAs, histone modifications, chromatin structures), proteomics (*in vitro* studies on proteins), and lipidomics (survey of cellular lipids), to name a few. The network medicine framework presents a promising way of thinking about and integrating these heterogeneous data types by elucidating their mutual influences to help explain disease etiologies and cellular functions and providing the basis for personalized therapeutics.

Multi-omics data integration using networks has already started gaining a wide amount of attention in the scientific community (Gligorijevic and Przulj, 2015; Tuncbag et al., 2016; Yugi et al., 2016; Hasin et al., 2017; Huang et al., 2017; Malod-Dognin et al., 2019). Moreover, relatively newer network tools like multiplex networks (Didier et al., 2018), network fusion (Wang et al., 2014), more innovative community detection strategies (Gligorijevic et al., 2016), and higher order structural modularity (Didier et al., 2018), have the potential to be applied to these problems to gain an even deeper and more nuance understanding of biological systems. Multilayer network approaches (De Domenico et al., 2015) for human diseases have unraveled important associations between rare and complex diseases (Halu et al., 2017). Despite several open challenges (Stegle et al., 2015; Ziegenhain et al., 2017), new technologies like single-cell transcriptomics (Hon et al., 2018), have started to be used to construct GRNs (Herbach et al., 2017; Fiers et al., 2018) and cell-specific coactivation networks (Ghazanfar et al., 2016). As the field of network medicine moves forward, one thing that is required more than ever before is the development of methods for systematically validating network predictions. Such validation will provide a greater confidence in network predictions and facilitate their incorporation into translational medicine. We also think active trans-disciplinary collaboration between biologists and scientists from the field of complex networks is required to infuse the field of network medicine with novel algorithms and innovative strategies.

The application of network methods to biomedical data presents a great opportunity to test and improve upon the tools originating from the general field of complex networks. We also take this opportunity to thank the many experimental biologists whose operose efforts have led to the generation of the vast amount of invaluable biomedical data, and to the numerous individuals who have donated their data for the sake of science.

AUTHOR CONTRIBUTIONS

ARS wrote the original draft which was reviewed, edited and revised by all the authors. All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

KG was supported by the NIH/NHLBI through K25HL133599. We acknowledge the support by National Institutes of Health (NIH) grants R01 HL118455-04-1 and P01 HL13285. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

AS would like to thank John Quackenbush for inspiration of the three paradigms discussed above, along with Trevor R. Leonardo and Rebekka Burkholz for critical reading of the manuscript. The authors thank members of the Quackenbush and Sharma labs for many fruitful discussions.

REFERENCES

Allen, N. E., Sudlow, C., Peakman, T., Collins, R., and Biobank, U. K. (2014). UK biobank data: come and get it. *Sci. Transl. Med.* 6:224ed4. doi: 10.1126/scitranslmed.3008601

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461. doi: 10.1038/nrg2102

Altaf-Ul-Amin, M., Afendi, F. M., Kiboi, S. K., and Kanaya, S. (2014). Systems biology in the context of big data and networks. *Biomed Res. Int.* 2014:428570. doi: 10.1155/2014/428570

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.* 9:e1002955. doi: 10.1371/journal.pcbi.1002955

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787

Aravind, L. (2000). Guilt by association: contextual information in genome analysis. *Genome Res.* 10, 1074–1077. doi: 10.1101/gr.10.8.1074

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M. K., Chuang, R., Jaehnig, E. J., et al. (2010). Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385–1389. doi: 10.1126/science.1195618

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Barrenas, F., Chavali, S., Alves, A. C., Coin, L., Jarvelin, M. R., Jornsten, R., et al. (2012). Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol.* 13:R46. doi: 10.1186/gb-2012-13-6-r46

Bauer, E., and Thiele, I. (2018). From network analysis to functional metabolic modeling of the human gut microbiota. *mSystems* 3:e00209-17. doi: 10.1128/mSystems.00209-17

Belanger, C. F., Hennekens, C. H., Rosner, B., and Speizer, F. E. (1978). The nurses' health study. *Am. J. Nurs.* 78, 1039–1040. doi: 10.2307/3462013

Bhardwaj, N., and Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics* 21, 2730–2738. doi: 10.1093/bioinformatics/bti398

Bouquet, J., Soloski, M. J., Swei, A., Cheadle, C., Federman, S., Billaud, J. N., et al. (2016). Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute lyme disease. *mBio* 7:e00100-16. doi: 10.1128/mBio.00100-16

Boyadjiev, S. A., and Jabs, E. W. (2000). Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* 57, 253–266. doi: 10.1034/j.1399-0004.2000.570403.x

Butte, A. J., and Kohane, I. S. (1999). Unsupervised knowledge discovery in medical databases using relevance networks. *Proc. AMIA Symp.* 1999, 711–715.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q., and Collins, J. J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903–915. doi: 10.1016/j.cell.2014.07.020

Cai, J. J., Borenstein, E., and Petrov, D. A. (2010). Broker genes in human disease. *Genome Biol. Evol.* 2, 815–825. doi: 10.1093/gbe/evq064

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Carter, H., Hofree, M., and Ideker, T. (2013). Genotype to phenotype via network analysis. *Curr. Opin. Genet. Dev.* 23, 611–621. doi: 10.1016/j.gde.2013.10.003

Celiku, O., Tandle, A., Chung, J. Y., Hewitt, S. M., Camphausen, K., and Shankavaram, U. (2017). Computational analysis of the mesenchymal signature landscape in gliomas. *BMC Med. Genomics* 10:13. doi: 10.1186/s12920-017-0252-7

Chen, C.-Y., Lopes-Ramos, C. M., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., et al. (2016). Sexual dimorphism in gene expression and regulatory networks across human tissues. *bioRxiv* [Preprint]. doi: 10.1101/082289

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307. doi: 10.1016/j.cell.2012.02.009

Chu, L. H., Rivera, C. G., Popel, A. S., and Bader, J. S. (2012). Constructing the angiome: a global angiogenesis protein interaction network. *Physiol. Genomics* 44, 915–924. doi: 10.1152/physiolgenomics.00181.2011

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Colditz, G. A., Philpott, S. E., and Hankinson, S. E. (2016). The impact of the nurses' health study on population health: prevention, translation, and control. *Am. J. Public Health* 106, 1540–1545. doi: 10.2105/AJPH.2016.303343

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science* 300, 286–290. doi: 10.1126/science.1084564

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: networks, competition, and stability. *Science* 350, 663–666. doi: 10.1126/science.aad2602

Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). Interactome: gateway into systems biology. *Hum. Mol. Genet.* 14(Suppl. 2), R171–R181. doi: 10.1093/hmg/ddi335

Davidson, E. H. (2006). *The Regulatory Genome : Gene Regulatory Networks in Development and Evolution*. Burlington, MA: Academic Press.

Davis, D., Yaveroglu, O. N., Malod-Dognin, N., Stojmirovic, A., and Przulj, N. (2015). Topology-function conservation in protein-protein interaction networks. *Bioinformatics* 31, 1632–1639. doi: 10.1093/bioinformatics/btv026

Dawber, T. R., Meadors, G. F., and Moore, F. E. Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. *Am. J. Public Health Nations Health* 41, 279–281. doi: 10.2105/AJPH.41.3.279

De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015). Structural reducibility of multilayer networks. *Nat. Commun.* 6:6864. doi: 10.1038/ncomms7864

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet.* 26, 326–333. doi: 10.1016/j.tig.2010.05.001

de Silva, E., and Stumpf, M. P. (2005). Complex networks and simple models in biology. *J. R. Soc. Interface* 2, 419–430. doi: 10.1098/rsif.2005.0067

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A. J. (2004). A gateway-compatible yeast one-hybrid system. *Genome Res.* 14, 2093–2101. doi: 10.1101/gr.2445504

Didier, G., Valdeolivas, A., and Baudot, A. (2018). Identifying communities from multiplex biological networks by randomized optimization of modularity. *F1000Res.* 7:1042. doi: 10.12688/f1000research.15486.2

Dondelinger, F., and Mukherjee, S. (2019). Statistical network inference for time-varying molecular data with dynamic bayesian networks. *Methods Mol. Biol.* 1883, 25–48. doi: 10.1007/978-1-4939-8882-2_2

Dreze, M., Monachello, D., Lurin, C., Cusick, M. E., Hill, D. E., Vidal, M., et al. (2010). High-quality binary interactome mapping. *Methods Enzymol.* 470, 281–315. doi: 10.1016/S0076-6879(10)70012-4

Driesbaugh, K. H., Branchetti, E., Grau, J. B., Keeney, S. J., Glass, K., Oyama, M. A., et al. (2018). Serotonin receptor 2B signaling with interstitial cell activation and leaflet remodeling in degenerative mitral regurgitation. *J. Mol. Cell. Cardiol.* 115, 94–103. doi: 10.1016/j.yjmcc.2017.12.014

Elo, L. L., Jarvenpaa, H., Oresic, M., Lahesmaa, R., and Aittokallio, T. (2007). Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. *Bioinformatics* 23, 2096–2103. doi: 10.1093/bioinformatics/btm309

Emamjomeh, A., Robat, E. S., Zahiri, J., Solouki, M., and Khosravi, P. (2017). Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnol. Rep.* 11, 71–86. doi: 10.1007/s11816-017-0433-z

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C. Y., Lopes-Ramos, C. M., et al. (2017). Exploring regulation in tissues with eQTL networks. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7841–E7850. doi: 10.1073/pnas.1707375114

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5:e8. doi: 10.1371/journal.pbio.0050008

Farahbod, M., and Pavlidis, P. (2018). Differential coexpression in human tissues and the confounding effect of mean expression levels. *Bioinformatics* 35, 55–61. doi: 10.1093/bioinformatics/bty538

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4323–4328. doi: 10.1073/pnas.0701722105

Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* 17, 246–254. doi: 10.1093/bfgp/elx046

Fournier, M. L., Paulson, A., Pavelka, N., Mosley, A. L., Gaudenz, K., Bradford, W. D., et al. (2010). Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin. *Mol. Cell Proteomics* 9, 271–284. doi: 10.1074/mcp.M900415-MCP200

Frey, B. J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315, 972–976. doi: 10.1126/science.1136800

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Ghazanfar, S., Bisogni, A. J., Ormerod, J. T., Lin, D. M., and Yang, J. Y. (2016). Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. *BMC Syst. Biol.* 10:127. doi: 10.1186/s12918-016-0370-4

Ghiassian, S. D., Menche, J., and Barabasi, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Ghiassian, S. D., Menche, J., Chasman, D. I., Giulianini, F., Wang, R., Ricchiuto, P., et al. (2016). Endophenotype network models: common core of complex diseases. *Sci. Rep.* 6:27414. doi: 10.1038/srep27414

Gillis, J., and Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* 8:e1002444. doi: 10.1371/journal.pcbi.1002444

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 898–907. doi: 10.1016/j.neuron.2011.05.021

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G. C. (2013). Passing messages between biological networks to refine predicted interactions. *PLoS One* 8:e64832. doi: 10.1371/journal.pone.0064832

Glass, K., Quackenbush, J., Silverman, E. K., Celli, B., Rennard, S. I., Yuan, G. C., et al. (2014). Sexually-dimorphic targeting of functionally-related genes in COPD. *BMC Syst. Biol.* 8:118. doi: 10.1186/s12918-014-0118-y

Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., and Yuan, G. C. (2015). A network model for angiogenesis in ovarian cancer. *BMC Bioinformatics* 16:115. doi: 10.1186/s12859-015-0551-y

Gligorijevic, V., Malod-Dognin, N., and Przulj, N. (2016). Fuse: multiple network alignment via data fusion. *Bioinformatics* 32, 1195–1203. doi: 10.1093/bioinformatics/btv731

Gligorijevic, V., and Przulj, N. (2015). Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface* 12:20150571. doi: 10.1098/rsif.2015.0571

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Goke, J., Jung, M., Behrens, S., Chavez, L., O'Keeffe, S., Timmermann, B., et al. (2011). Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput. Biol.* 7:e1002304. doi: 10.1371/journal.pcbi.1002304

Grobbee, D. E., Rimm, E. B., Giovannucci, E., Colditz, G., Stampfer, M., and Willett, W. (1990). Coffee, caffeine, and cardiovascular disease in men. *N. Engl. J. Med.* 323, 1026–1032. doi: 10.1056/NEJM199010113231504

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Guney, E., Menche, J., Vidal, M., and Barabasi, A. L. (2016). Network-based in silico drug efficacy screening. *Nat. Commun.* 7:10331. doi: 10.1038/ncomms10331

Guo, Y., Alexander, K., Clark, A. G., Grimson, A., and Yu, H. (2016). Integrated network analysis reveals distinct regulatory roles of transcription factors and microRNAs. *RNA* 22, 1663–1672. doi: 10.1261/rna.048025.114

Hall-Swan, S., Crawford, J., Newman, R., and Cowen, L. J. (2018). Detangling PPI networks to uncover functionally meaningful clusters. *BMC Syst. Biol.* 12:24. doi: 10.1186/s12918-018-0550-5

Halu, A., De Domenico, M., Arenas, A., and Sharma, A. (2017). The multiplex network of human diseases. *bioRxiv* [Preprint]. doi: 10.1101/100370

Halu, A., Wang, J. G., Iwata, H., Mojcher, A., Abib, A. L., Singh, S. A., et al. (2018). Context-enriched interactome powered by proteomics helps the identification of novel regulators of macrophage activation. *eLife* 7:e37059. doi: 10.7554/eLife.37059

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52–55. doi: 10.1093/nar/30.1.52

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol.* 7:120.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1

He, B., and Tan, K. (2016). Understanding transcriptional regulatory networks using computational models. *Curr. Opin. Genet. Dev.* 37, 101–108. doi: 10.1016/j.gde.2016.02.002

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., et al. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.* 43, 414–420. doi: 10.1038/ng.808

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163, 712–723. doi: 10.1016/j.cell.2015.09.053

Hemberg, M., and Kreiman, G. (2011). Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res.* 39, 7092–7102. doi: 10.1093/nar/gkr404

Herbach, U., Bonnaffoux, A., Espinasse, T., and Gandrillon, O. (2017). Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol.* 11:105. doi: 10.1186/s12918-017-0487-0

Hochedlinger, K., and Plath, K. (2009). Epigenetic reprogramming and induced pluripotency. *Development* 136, 509–523. doi: 10.1242/dev.020867

Hon, C. C., Shin, J. W., Carninci, P., and Stubbington, M. J. T. (2018). The Human Cell Atlas: technical approaches and challenges. *Brief. Funct. Genomics* 17, 283–294. doi: 10.1093/bfgp/elx029

Huang, J., Zheng, J., Yuan, H., and McGinnis, K. (2018). Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biol.* 18:111. doi: 10.1186/s12870-018-1329-y

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., et al. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* 6, 484–495.e5. doi: 10.1016/j.cels.2018.03.001

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509. doi: 10.1038/nature22366

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., et al. (2015). The BioPlex network: a systematic exploration of the human interactome. *Cell* 162, 425–440. doi: 10.1016/j.cell.2015.06.043

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8:565. doi: 10.1038/msb.2011.99

Ideker, T., and Sharan, R. (2008). Protein networks in disease. *Genome Res.* 18, 644–652. doi: 10.1101/gr.071852.107

Jaini, S., Lyubetskaya, A., Gomes, A., Peterson, M., Tae Park, S., Raman, S., et al. (2014). Transcription factor binding site mapping using ChIP-Seq. *Microbiol. Spectr.* 2:MGM2-MGM0035. doi: 10.1128/microbiolspec.MGM2-0035-2013

Jalan, S., Solymosi, N., Vattay, G., and Li, B. (2010). Random matrix analysis of localization properties of gene coexpression network. *Phys. Rev. E* 81(4 Pt 2):046118. doi: 10.1103/PhysRevE.81.046118

Jalan, S., Ung, C. Y., Bhojwani, J., Li, B., Zhang, L., Lan, S. H., et al. (2012). Spectral analysis of gene co-expression network of Zebrafish. *EPL* 99:48004. doi: 10.1209/0295-5075/99/48004

Jambusaria, A., Klomp, J., Hong, Z., Rafii, S., Dai, Y., Malik, A. B., et al. (2018). A computational approach to identify cellular heterogeneity and tissue-specific gene regulatory networks. *BMC Bioinformatics* 19:217. doi: 10.1186/s12859-018-2190-6

Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297. doi: 10.1093/bioinformatics/btl390

Jung, M., Jin, S. G., Zhang, X., Xiong, W., Gogoshin, G., Rodin, A. S., et al. (2015). Longitudinal epigenetic and gene expression profiles analyzed by three-component analysis reveal down-regulation of genes involved in protein translation in human aging. *Nucleic Acids Res.* 43:e100. doi: 10.1093/nar/gkv473

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37, D623–D628. doi: 10.1093/nar/gkn698

Kanaya, S., Altaf-Ul-Amin, M., Kiboi, S. K., and Mochamad Afendi, F. (2014). Big data and network biology. *Biomed Res. Int.* 2014:836708. doi: 10.1155/2014/836708

Kanonidis, E. I., Roy, M. M., Deighton, R. F., and Le Bihan, T. (2016). Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: case of a glioblastoma multiforme study. *PLoS One* 11:e0161828. doi: 10.1371/journal.pone.0161828

Kim, R. S., Ji, H., and Wong, W. H. (2006). An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse. *BMC Bioinformatics* 7:44. doi: 10.1186/1471-2105-7-44

Kim, Y., Han, S., Choi, S., and Hwang, D. (2014). Inference of dynamic networks using time-course data. *Brief. Bioinform.* 15, 212–228. doi: 10.1093/bib/bbt028

Kim, Y., and Kim, J. (2018). Estimation of dynamic systems for gene regulatory networks from dependent time-course data. *J. Comput. Biol.* 25, 987–996. doi: 10.1089/cmb.2018.0062

Kitsak, M., Sharma, A., Menche, J., Guney, E., Ghiassian, S. D., Loscalzo, J., et al. (2016). Tissue specificity of human disease module. *Sci. Rep.* 6:35241. doi: 10.1038/srep35241

Klingstrom, T., and Plewczynski, D. (2011). Protein-protein interaction and pathway databases, a graphical review. *Brief. Bioinform.* 12, 702–713. doi: 10.1093/bib/bbq064

Kogelman, L. J., Fu, J., Franke, L., Greve, J. W., Hofker, M., Rensen, S. S., et al. (2016). Inter-tissue gene co-expression networks between metabolically healthy and unhealthy obese individuals. *PLoS One* 11:e0167519. doi: 10.1371/journal.pone.0167519

Lage, K., Hansen, N. T., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20870–20875. doi: 10.1073/pnas.0810772105

Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316. doi: 10.1038/nbt1295

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.* 25, 217–228. doi: 10.1016/j.tim.2016.11.008

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21, 1109–1121. doi: 10.1101/gr.118992.110

Li, P., Zhang, C., Perkins, E. J., Gong, P., and Deng, Y. (2007). Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics* 8(Suppl. 7):S13. doi: 10.1186/1471-2105-8-S7-S13

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., et al. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* 303, 540–543.

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liesecke, F., Daudu, D., Duge de Bernonville, R., Besseau, S., Clastre, M., Courdavault, V., et al. (2018). Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* 8:10885. doi: 10.1038/s41598-018-29077-3

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J.-F., et al. (2006). A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125, 801–814. doi: 10.1016/j.cell.2006.03.032

Lin, J. S., and Lai, E. M. (2017). Protein-protein interactions: co-immunoprecipitation. *Methods Mol. Biol.* 1615, 211–219. doi: 10.1007/978-1-4939-7033-9_17

Lindfors, E. (2011). *Network Biology: Applications in Medicine and Biotechnology*. Espoo: VTT.

Liu, E. T., and Lauffenburger, D. A. (2009). *Systems Biomedicine: Concepts and Perspectives* 1 Amsterdam: Elsevier.

Liu, W., Ma, L., Jeon, B., Chen, L., and Chen, B. (2018). A Network Hierarchy-Based method for functional module detection in protein-protein interaction networks. *J. Theor. Biol.* 455, 26–38. doi: 10.1016/j.jtbi.2018.06.026

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell* 165, 535–550. doi: 10.1016/j.cell.2016.03.014

Liu, Y. Y., Slotine, J. J., and Barabasi, A. L. (2011). Controllability of complex networks. *Nature* 473, 167–173. doi: 10.1038/nature10011

Lopes-Ramos, C. M., Kuijjer, M. L., Ogino, S., Fuchs, C. S., DeMeo, D. L., Glass, K., et al. (2018). Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer Res.* 78, 5538–5547. doi: 10.1158/0008-5472.CAN-18-0454

Lopes-Ramos, C. M., Paulson, J. N., Chen, C. Y., Kuijjer, M. L., Fagny, M., Platig, J., et al. (2017). Regulatory network changes between cell lines and their tissues of origin. *BMC Genomics* 18:723. doi: 10.1186/s12864-017-4111-x

Loscalzo, J., Barabaìsi, A. -L. S., and Silverman, E. K. (2017). *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Cambridge, MA: Harvard University Press.

Lu, H. C., Fornili, A., and Fraternali, F. (2013). Protein-protein interaction networks studies and importance of 3D structure knowledge. *Expert Rev. Proteomics* 10, 511–520. doi: 10.1586/14789450.2013.856764

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K., et al. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* 8:299. doi: 10.1186/1471-2105-8-299

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8:573. doi: 10.1038/s41467-017-00680-8

Mahmood, S. S., Levy, D., Vasan, R. S., and Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* 383, 999–1008. doi: 10.1016/S0140-6736(13)61752-3

Malod-Dognin, N., Ban, K., and Przulj, N. (2017). Unified alignment of protein-protein interaction networks. *Sci. Rep.* 7:953. doi: 10.1038/s41598-017-01085-9

Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemmingway, H., Ketteler, R., et al. (2019). Towards a data–integrated cell. *Nat. Commun.* 10:805. doi: 10.1038/s41467-019-08797-8

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Marco, A., Konikoff, C., Karr, T. L., and Kumar, S. (2009). Relationship between gene co-expression and sharing of transcription factor binding sites in *Drosophila melanogaster*. *Bioinformatics* 25, 2473–2477. doi: 10.1093/bioinformatics/btp462

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bahler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 151, 671–683. doi: 10.1016/j.cell.2012.09.019

McDermott, J. E., Diamond, D. L., Corley, C., Rasmussen, A. L., Katze, M. G., and Waters, K. M. (2012). Topological analysis of protein co-abundance networks identifies novel host targets important for HCV infection and pathogenesis. *BMC Syst. Biol.* 6:28. doi: 10.1186/1752-0509-6-28

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol.* 2007:79879. doi: 10.1155/2007/79879

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824–827. doi: 10.1126/science.298.5594.824

Min, L., Zhang, C., Qu, L., Huang, J., Jiang, L., Liu, J., et al. (2017). Gene regulatory pattern analysis reveals essential role of core transcriptional factors' activation in triple-negative breast cancer. *Oncotarget* 8, 21938–21953. doi: 10.18632/oncotarget.15749

Mordelet, F., and Vert, J. P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24, i76–i82. doi: 10.1093/bioinformatics/btn273

Moreau, Y., and Tranchevent, L. C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* 13, 523–536. doi: 10.1038/nrg3253

Mosca, R., Pons, T., Ceol, A., Valencia, A., and Aloy, P. (2013). Towards a detailed atlas of protein-protein interactions. *Curr. Opin. Struct. Biol.* 23, 929–940. doi: 10.1016/j.sbi.2013.07.005

Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., and Lu, T. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* 13, 2847–2852. doi: 10.4161/15384101.2014.949201

Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26, 1057–1063. doi: 10.1093/bioinformatics/btq076

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69:066133. doi: 10.1103/PhysRevE.69.066133

Ni, J., Koyuturk, M., Tong, H., Haines, J., Xu, R., and Zhang, X. (2016). Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. *BMC Bioinformatics* 17:453. doi: 10.1186/s12859-016-1317-x

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J. Med. Genet.* 43, 691–698. doi: 10.1136/jmg.2006.041376

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., and Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. *Methods* 74, 47–53. doi: 10.1016/j.ymeth.2014.10.026

Perrin, B. E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alche-Buc, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19(Suppl. 2), ii138–ii148. doi: 10.1093/bioinformatics/btg1071

Pierson, E., Consortium, G. T., Koller, D., Battle, A., Mostafavi, S., Ardlie, K. G., et al. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.* 11:e1004220. doi: 10.1371/journal.pcbi.1004220

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., et al. (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* 5:e1000421. doi: 10.1371/journal.pgen.1000421

Pinero, J., Berenstein, A., Gonzalez-Perez, A., Chernomoretz, A., and Furlong, L. I. (2016). Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Sci. Rep.* 6:24570. doi: 10.1038/srep24570

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Pirgazi, J., and Khanteymoori, A. R. (2018). A robust gene regulatory network inference method base on Kalman filter and linear regression. *PLoS One* 13:e0200094. doi: 10.1371/journal.pone.0200094

Platig, J., Castaldi, P. J., DeMeo, D., and Quackenbush, J. (2016). Bipartite community structure of eQTLs. *PLoS Comput. Biol.* 12:e1005033. doi: 10.1371/journal.pcbi.1005033

Prieto, C., Risueno, A., Fontanillo, C., and De las Rivas, J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One* 3:e3911. doi: 10.1371/journal.pone.0003911

Prokopenko, D., Sakornsakolpat, P., Fier, H. L., Qiao, D., Parker, M. M., McDonald, M. N., et al. (2018). Whole-genome sequencing in severe chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* 59, 614–622. doi: 10.1165/rcmb.2018-0088OC

Przulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics* 23, e177–e183. doi: 10.1093/bioinformatics/btl301

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 3508–3515.

Qiu, W., Guo, F., Glass, K., Yuan, G. C., Quackenbush, J., Zhou, X., et al. (2018). Differential connectivity of gene regulatory networks distinguishes corticosteroid response in asthma. *J. Allergy Clin. Immunol.* 141, 1250–1258. doi: 10.1016/j.jaci.2017.05.052

Quackenbush, J. (2003). Genomics. Microarrays–guilt by association. *Science* 302, 240–241. doi: 10.1126/science.1090887

Rackham, O. J., Firas, J., Fang, H., Oates, M. E., Holmes, M. L., Knaupp, A. S., et al. (2016). A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* 48, 331–335. doi: 10.1038/ng.3487

Reverter, A., and Chan, E. K. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24, 2491–2497. doi: 10.1093/bioinformatics/btn482

Rodenas-Cuadrado, P., Chen, X. S., Wiegrebe, L., Firzlaff, U., and Vernes, S. C. (2015). A novel approach identifies the first transcriptome networks in bats: a new genetic model for vocal communication. *BMC Genomics* 16:836. doi: 10.1186/s12864-015-2068-1

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L., and Almouzni, G. (2015). Epigenomics: roadmap for regulation. *Nature* 518, 314–316. doi: 10.1038/518314a

Rotival, M., and Petretto, E. (2014). Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits. *Brief. Funct. Genomics* 13, 66–78. doi: 10.1093/bfgp/elt030

Rottjers, L., and Faust, K. (2018). From hairballs to hypotheses-biological insights from microbial networks. *FEMS Microbiol. Rev.* 42, 761–780. doi: 10.1093/femsre/fuy030

Roy, S., Bhattacharyya, D. K., and Kalita, J. K. (2014). Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics* 15(Suppl. 7):S10. doi: 10.1186/1471-2105-15-S7-S10

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., and Teichmann, S. A. (2017). The Human Cell Atlas: from vision to reality. *Nature* 550, 451–453. doi: 10.1038/550451a

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178. doi: 10.1038/nature04209

Ryan, C. J., Kennedy, S., Bajrami, I., Matallanas, D., and Lord, C. J. (2017). A compendium of co-regulated protein complexes in breast cancer reveals collateral loss events. *Cell Syst.* 5, 399–409.e5. doi: 10.1016/j.cels.2017.09.011

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., et al. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* 27, 1843–1858. doi: 10.1101/gr.216721.116

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647–660. doi: 10.1016/j.cell.2015.04.013

Schadt, E. E., and Bjorkegren, J. L. (2012). NEW: network-enabled wisdom in biology, medicine, and health care. *Sci. Transl. Med.* 4:115rv1. doi: 10.1126/scitranslmed.3002132

Schweppe, D. K., Huttlin, E. L., Harper, J. W., and Gygi, S. P. (2018). BioPlex display: an interactive suite for large-scale AP-MS protein-protein interaction data. *J. Proteome Res.* 17, 722–726. doi: 10.1021/acs.jproteome.7b00572

Serin, E. A., Nijveen, H., Hilhorst, H. W., and Ligterink, W. (2016). Learning from Co-expression Networks: possibilities and challenges. *Front. Plant Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Sharma, A., Chavali, S., Tabassum, R., Tandon, N., and Bharadwaj, D. (2010). Gene prioritization in Type 2 Diabetes using domain interactions and network analysis. *BMC Genomics* 11:84. doi: 10.1186/1471-2164-11-84

Sharma, A., Gulbahce, N., Pevzner, S. J., Menche, J., Ladenvall, C., Folkersen, L., et al. (2013). Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. *Mol. Cell. Proteomics* 12, 3398–3408. doi: 10.1074/mcp.M112.024851

Sharma, A., Kitsak, M., Cho, M. H., Ameli, A., Zhou, X., Jiang, Z., et al. (2018). Integration of molecular interactome and targeted interaction analysis to identify a COPD disease network module. *Sci. Rep.* 8:14439. doi: 10.1038/s41598-018-32173-z

Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* 24, 3005–3020. doi: 10.1093/hmg/ddv001

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U.S.A.* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68. doi: 10.1038/ng881

Sonawane, A. R., Platig, J., Fagny, M., Chen, C. Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi: 10.1016/j.celrep.2017.10.001

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13:328. doi: 10.1186/1471-2105-13-328

Spirin, V., and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 12123–12128. doi: 10.1073/pnas.2032324100

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255. doi: 10.1126/science.1087447

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12:e1001779. doi: 10.1371/journal.pmed.1001779

Sun, P. G., Gao, L., and Han, S. (2011). Prediction of human disease-related gene clusters by clustering analysis. *Int. J. Biol. Sci.* 7, 61–73. doi: 10.7150/ijbs.7.61

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., et al. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* 27, 199–204. doi: 10.1038/nbt.1522

Terzer, M., Maynard, N. D., Covert, M. W., and Stelling, J. (2009). Genome-scale metabolic networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 1, 285–297. doi: 10.1002/wsbm.37

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

The International HapMap Consortium (2003). The international HapMap project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Tian, Z., Guo, M., Wang, C., Xing, L., Wang, L., and Zhang, Y. (2017). Constructing an integrated gene similarity network for the identification of disease genes. *J. Biomed. Semantics* 8(Suppl. 1):32. doi: 10.1186/s13326-017-0141-1

Toh, H., and Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* 18, 287–297. doi: 10.1093/bioinformatics/18.2.287

Tuncbag, N., Gosline, S. J., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLoS Comput. Biol.* 12:e1004879. doi: 10.1371/journal.pcbi.1004879

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627. doi: 10.1038/35001009

Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A., Amini, B., Andersen, E., et al. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 4, 1920–1932. doi: 10.1074/mcp.M500279-MCP200

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

van Someren, E. P., Vaes, B. L., Steegenga, W. T., Sijbers, A. M., Dechering, K. J., and Reinders, M. J. (2006). Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics* 22, 477–484. doi: 10.1093/bioinformatics/bti816

Vargas, A. J., Quackenbush, J., and Glass, K. (2016). Diet-induced weight loss leads to a switch in gene regulatory network control in the rectal mucosa. *Genomics* 108, 126–133. doi: 10.1016/j.ygeno.2016.08.001

Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., et al. (2009). An empirical framework for binary interactome mapping. *Nat. Methods* 6, 83–90. doi: 10.1038/nmeth.1280

Vidal, M., Cusick, M. E., and Barabasi, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Vidal, M., and Fields, S. (2014). The yeast two-hybrid assay: still finding connections after 25 years. *Nat. Methods* 11, 1203–1206. doi: 10.1038/nmeth.3182

Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., et al. (2018). ModuleDiscoverer: identification of regulatory modules in protein-protein interaction networks. *Sci. Rep.* 8:433. doi: 10.1038/s41598-017-18370-2

Vogel, C., Silva, G. M., and Marcotte, E. M. (2011). Protein expression regulation under oxidative stress. *Mol. Cell. Proteomics* 10:M111.009217. doi: 10.1074/mcp.M111.009217

Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21, 4205–4208. doi: 10.1093/bioinformatics/bti688

Walhout, M., Vidal, M., and Dekker, J. (2012). *Handbook of Systems Biology: Concepts and Insights*. Oxford: Academic Press.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, Q., Tan, Y. X., Ren, Y. B., Dong, L. W., Xie, Z. F., Tang, L., et al. (2011). Zinc finger protein ZBTB20 expression is increased in hepatocellular carcinoma and associated with poor prognosis. *BMC Cancer* 11:271. doi: 10.1186/1471-2407-11-271

Wang, S., Ma, J., Zhang, W., Shen, J. P., Huang, J., Peng, J., et al. (2018). Typing tumors using pathways selected by somatic evolution. *Nat. Commun.* 9:4159. doi: 10.1038/s41467-018-06464-y

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics* 22, 2523–2531. doi: 10.1093/bioinformatics/btl391

Will, T., and Helms, V. (2014). Identifying transcription factor complexes and their roles. *Bioinformatics* 30, i415–i421. doi: 10.1093/bioinformatics/btu448

Willsey, A. J., Morris, M. T., Wang, S., Willsey, H. R., Sun, N., Teerikorpi, N., et al. (2018). The psychiatric cell map initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders. *Cell* 174, 505–520. doi: 10.1016/j.cell.2018.06.016

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schutte, J., Kaimakis, P., et al. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* 7, 532–544. doi: 10.1016/j.stem.2010.07.016

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.* 4:189. doi: 10.1038/msb.2008.27

Wysocki, K., and Ritter, L. (2011). Diseasome: an approach to understanding gene-disease interactions. *Annu. Rev. Nurs. Res.* 29, 55–72. doi: 10.1891/0739-6686.29.55

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291. doi: 10.1093/nar/28.1.289

Xiao, X., Moreno-Moral, A., Rotival, M., Bottolo, L., and Petretto, E. (2014). Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genet.* 10:e1004006. doi: 10.1371/journal.pgen.1004006

Xing, L., Guo, M., Liu, X., Wang, C., Wang, L., and Zhang, Y. (2017). An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. *BMC Genomics* 18(Suppl. 9):844. doi: 10.1186/s12864-017-4228-y

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* 5:3231. doi: 10.1038/ncomms4231

Yeung, M. K., Tegner, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6163–6168. doi: 10.1073/pnas.092576199

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.* 25, 1119–1126. doi: 10.1038/nbt1338

Yu, H., Luscombe, N. M., Qian, J., and Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* 19, 422–427. doi: 10.1016/S0168-9525(03)00175-6

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355–364. doi: 10.1038/nature13992

Yugi, K., Kubota, H., Hatano, A., and Kuroda, S. (2016). Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers. *Trends Biotechnol.* 34, 276–290. doi: 10.1016/j.tibtech.2015.12.013

Zhang, Q. C., Petrey, D., Garzon, J. I., Deng, L., and Honig, B. (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.* 41, D828–D833. doi: 10.1093/nar/gks1231

Zhang, X., Zhang, R., Jiang, Y., Sun, P., Tang, G., Wang, X., et al. (2011). The expanded human disease network combining protein-protein interaction information. *Eur. J. Hum. Genet.* 19, 783–788. doi: 10.1038/ejhg.2011.30

Zhou, X., Lei, L., Liu, J., Halu, A., Zhang, Y., Li, B., et al. (2018). A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *EBioMedicine* 31, 79–91. doi: 10.1016/j.ebiom.2018.04.002

Zhou, X., Menche, J., Barabasi, A. L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643.e4. doi: 10.1016/j.molcel.2017.01.023

Zuo, Y., Cui, Y., Di Poto, C., Varghese, R. S., Yu, G., Li, R., et al. (2016). INDEED: integrated differential expression and differential network analysis of omic data for biomarker discovery. *Methods* 111, 12–20. doi: 10.1016/j.ymeth.2016.08.015

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.

Copyright © 2019 Sonawane, Weiss, Glass and Sharma. 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) and the copyright owner(s) 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 Genetics

METHODS
published: 22 January 2019
doi: 10.3389/fgene.2019.00004

Comparative Analysis of Normalization Methods for Network Propagation

*Hadas Biran[1], Martin Kupiec[2] and Roded Sharan[3]**

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel, [2] School of Molecular Cell Biology and Biotechnology, Tel Aviv University, Tel Aviv, Israel, [3] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Network propagation is a central tool in biological research. While a number of variants and normalizations have been proposed for this method, each has its own shortcomings and no large scale assessment of those variants is available. Here we propose a novel normalization method for network propagation that is based on evaluating the propagation results against those obtained on randomized networks that preserve node degrees. In this way, our method overcomes potential biases of previous methods. We evaluate its performance on multiple large scale datasets and find that it compares favorably to previous approaches in diverse gene prioritization tasks. We further demonstrate its utility on a focused dataset of telomere length maintenance in yeast. The normalization method is available at http://anat.cs.tau.ac.il/WebPropagate.

Keywords: network diffusion, protein–protein interaction network, gene prioritization, *p*-value computation, degree-preserving randomization, telomere length maintenance

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Mehmet Koyuturk,
Case Western Reserve University, United States
Evan Oliver Paull,
Columbia University, United States

***Correspondence:**
Roded Sharan
roded@tau.ac.il

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 October 2018*
Accepted: *07 January 2019*
Published: *22 January 2019*

Citation:
Biran H, Kupiec M and Sharan R (2019) Comparative Analysis of Normalization Methods for Network Propagation. Front. Genet. 10:4. doi: 10.3389/fgene.2019.00004

INTRODUCTION

Network propagation is a method of choice for diverse analyses such as protein function prediction, gene prioritization and identification of disease modules (Cowen et al., 2017). There are at least 17 available software tools that employ different variants of network propagation for these purposes (Cowen et al., 2017; Biran et al., 2018).

However, the basic propagation technique has some known limitations: First, raw propagation scores do not carry any statistical significance information and can only be used to rank proteins. Second, they are greatly affected by the degrees of initial proteins implicated in the process under study (termed seed set below) and the degree of any candidate protein being scored. This biases the results toward high degree, well studied proteins.

To deal with the second challenge, Erten et al. (2011) suggested the DADA normalization approach. This method normalizes the raw propagation scores with the eigenvector centrality measure for each protein, and then produces ranks based on either these normalizations or the raw propagation scores, depending on the seed set average weighted degree.

Mazza et al. (2016) tackled the first challenge by evaluating propagation scores against those obtained from propagating random seed sets. Nevertheless, none of the methods solves both problems, calling for a more complete solution.

In this work we present a novel normalization technique that tackles both challenges. We developed a new technique, in which the raw propagation scores are normalized through propagation scores obtained in random degree-preserving networks (RDPN). In cross validation tests, our method outperforms previous normalizations in gene prioritization tasks on diverse disease-related and function-related data sets in both human and yeast. Furthermore, it eliminates the degree biases of previous approaches and allows the assessment of statistical significance of the results by providing *p*-values that are corrected for multiple testing of candidate proteins.

RESULTS

Network Propagation

Network propagation is a process in which a preselected set of seed proteins that underlie some phenotype of interest are viewed as "heat sources" in a PPI network. The heat is diffused to the rest of the proteins in the network in an iterative process until a steady-state is attained. Proteins that are relatively close to the seed set get higher propagation scores than distant proteins and are therefore considered to be associated with the phenotype in question. Network propagation is widely used for protein prioritization and related tasks (Cowen et al., 2017).

Formally, given a binary vector P_0 denoting seed proteins, a normalized network adjacency matrix W (see below) and a smoothing parameter α controlling the relative importance of the network vs. the seed information, it can be shown that the propagation process converges to a score vector.

$$P = (1 - a)\,(I - \alpha W)^{-1}\,P_0$$

Henceforth, we follow (Vanunu et al., 2010) and set $\alpha = 0.8$ (unless stated otherwise), to allow a fairly high network influence over the prior (seed) knowledge.

There are two main ways by which the adjacency matrix A (which could be weighted or unweighted) is normalized to ensure the convergence of the process: (i) a symmetric variant in, which $W = D^{-1/2}AD^{-1/2}$ and (ii) a degree-based variant, in which $W = AD^{-1}$. Here D denotes the diagonal weighted degree matrix.

Previously Suggested Normalization Solutions

The raw scores from the propagation process do not carry a statistical meaning, and highly depend on the size of the seed set and the degrees of the proteins involved. It is thus desirable to normalize them. In the following we describe three previous normalization methods and a new hybrid of two of the methods; full details can be found in the Methods.

Erten et al. (2011) suggested the DADA method that builds on normalizing each propagation score by the eigenvector centrality measure of the same protein, which can be calculated by propagating with $\alpha = 1$ from the same seed set (Brin and Page, 1998; Bryan and Leise, 2006; Erten et al., 2011). Here we analyze both this simple EC method and the full DADA method which uses ranks (rather than the scores themselves) of the regular propagation scores in case the average weighted degree of the seed set exceeds the network average weighted degree, or the logarithm of the EC score otherwise.

Mazza et al. (2016) suggested normalizing propagation scores by comparing them to propagations from random seed sets (RSS). This method produces *p*-values and is implemented as a web tool at http://anat.cs.tau.ac.il/WebPropagate/ (Biran et al., 2018).

We also examine here a hybrid of RSS and DADA, which we call RSS_SD. This variant produces *p*-values in the same manner RSS does, but the random seed sets are chosen to be degree-distributed like the original seed set using the method of Erten et al. (2011).

Normalization With Random Degree-Preserving Networks (RDPN)

The only previous normalization method we are aware of that assigns statistical significance to the propagation scores is based on propagating random seed sets. Such computations do not take into account the degrees of the seed nodes. To overcome this shortcoming, we propose a novel method that is based on randomizations of the input network rather than the seed sets. Specifically, the propagation score of a protein is compared to the scores the protein attains on random degree-preserving networks under the same seed set. Our normalization method with random degree-preserving networks, RDPN, is schematically depicted in **Figure 1**.

In order to execute this method, one first has to compute n random degree-preserving networks (we use $n = 100$ unless otherwise stated). We implemented the "switching" method, in which in each iteration two edges (u, v) and (s, t) are picked randomly, and if $u \neq v \neq s \neq t$ and the edges (u, t), (s, v) do not already exist, then they are "switched," namely the edges (u, v) and (s, t) are removed and the edges (u, t) and (s, v) are added. For the construction of one random network, we executed $100^*|E|$ such iterations, where $|E|$ denotes the number of edges in the network, per the recommendation in Milo et al. (2003).

One issue that immediately emerges is the question of connectivity. Network propagation relies on the fact that all relevant proteins are part of one connected component, otherwise the information will not diffuse in a desired way. For example, suppose that during the randomization process two proteins got disconnected from the main component, creating a very small connected component of their own. If one of them is a seed protein, then the propagation score of the other one will be unreasonably high. However, if none of them is a seed protein, then their propagation scores will be 0. We addressed this issue by considering for each protein only the instances in which it was part of the main connected component in the network.

In detail, *p*-values are computed as follows: Each protein v gets a "real" propagation score X^v_{real} by propagating from the seed set on the original network; it also gets n random scores X^v_i

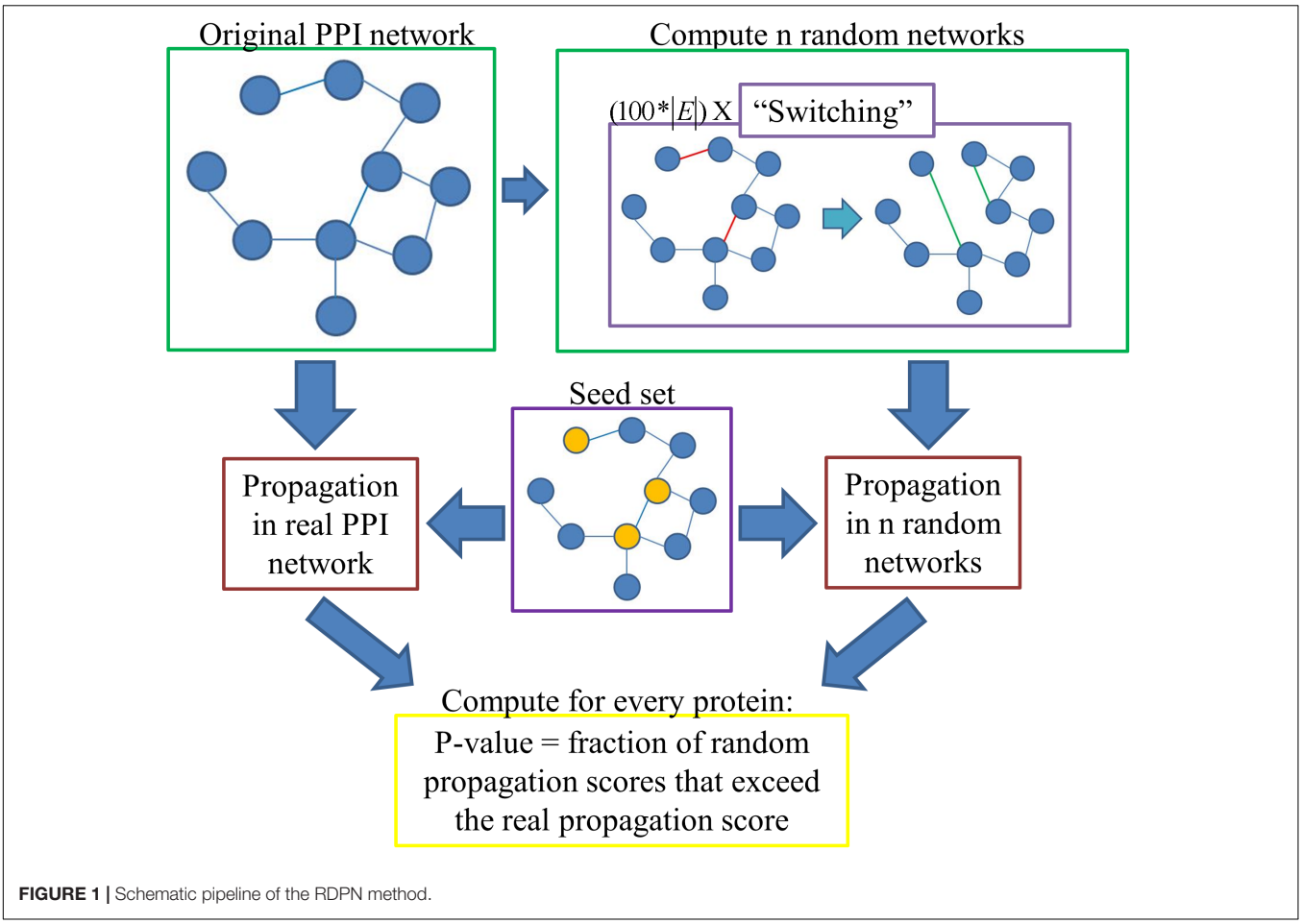


FIGURE 1 | Schematic pipeline of the RDPN method.

$(0 \leq i \leq n-1)$ by propagating from the same seed set on the n random networks. Then its p-value is computed as the fraction of random instances in which its score exceeded its real propagation score, i.e.:

$$p^v = \frac{|\{i|(X_i^v \geq X_{real}^v \text{ and } v \text{ is part of the main connected component in the } i\text{'th network})\}| + 1}{|\{i|(v \text{ is part of the main connected component in the } i\text{'th network})\}| + 1}$$

To overcome the infrequent case in which a protein has a high tendency to get disconnected and, therefore, its p-value is determined based on an insufficient number of instances, we determined that a protein with less than $n/2$ relevant instances (instances in which it was part of the main connected component) will be assigned a p-value of one. Empirically, in our pre-computed random networks there was no such protein and therefore this condition was never used.

Performance Evaluation

We compared the basic propagation computation with the three previously suggested normalization techniques (EC, DADA, and RSS), RSS_SD and our own Random Degree-Preserving Networks (RDPN) normalization with respect to their performance in multiple disease-related and function-related prioritization tasks as described below.

Overall Performance

We evaluated the performance of the six methods and two matrix normalization variants on four large-scale data sets in a fivefold cross validation setting. Each data set contained multiple groups of function-related or disease-related genes with respect to which the prioritization of each normalization method was evaluated. Each method's performance was summarized by the area under the ROC curve (AUROC) measure, when using similar-degree negative samples (Methods).

The evaluation results are given in **Table 1**. Regarding the two variants of adjacency matrix normalization, we found that in 12 out of 24 method-data set pairs (and also on average) the symmetric variant performs better (in 10 of them the degree-based variant performed better, and 2 were ties). Therefore, we focused on this variant in all subsequent evaluations. On average, the three top performing normalization methods were RDPN, RSS_SD, and EC, attaining similar AUROCs across the four data sets.

However, when examining the performance on the individual groups within the data sets, we found that the RDPN method greatly outperformed all others with the highest number of

TABLE 1 | Average AUROC of the six methods across four data sets, using two variants of adjacency matrix normalization.

Dataset	Symmetric adjacency matrix normalization						Degree-based adjacency matrix normalization					
	Propagation	EC	DADA	RSS	RSS_SD	RDPN	Propagation	EC	DADA	RSS	RSS_SD	RDPN
Menche-OMIM	0.695	0.74	0.707	0.729	0.745	**0.746**	0.663	**0.742**	0.685	0.738	**0.742**	**0.742**
GO_MF	0.76	0.83	0.783	0.805	0.827	**0.832**	0.715	0.83	0.749	0.826	**0.832**	0.831
GO_CC	0.763	**0.833**	0.782	0.812	0.829	**0.833**	0.721	**0.833**	0.75	0.83	**0.833**	0.831
GO_BP	0.74	0.798	0.757	0.774	0.797	**0.801**	0.707	0.802	0.734	0.798	0.8	**0.803**

For each dataset, the best performing method in each variant is shown in bold.

groups for which it gave the best results across all data sets (**Figure 2**).

Degree Bias of the Different Methods

A good normalization method should account for the degrees of the candidate proteins, as these influence propagation scores. To test this, we focused on the Menche-OMIM set. Expectedly, the raw propagation scores are highly correlated with the weighted degree of the candidate protein (0.901 Spearman correlation). A similar anti-correlation level (−0.749) was observed for DADA's ranks. In contrast, EC scores were only weakly correlated with the candidate protein weighted degree (average Spearman coefficient of 0.238), and the *p*-values computed by RSS, RSS_SD, and RDPN were relatively unbiased (average Spearman coefficients of 0.019, 0.035, and 0.078, respectively). These results are depicted in **Figure 3**.

P-Value Biases

While the regular propagation, EC and DADA produce scores or ranks, which are only expected to be meaningful for ranking proteins within the same run, RSS, RSS_SD, and RDPN produce *p*-values, which can be thresholded within and across runs to yield statistically significant hits. In order to evaluate the robustness of the assigned *p*-values, we tested their dependence on the average weighted degree of the seed set, focusing on the Menche-OMIM set. We found that both RDPN's and RSS_SD's percents of significant hits (p-value < 0.05) are only mildly affected by the seed set average weighted degree (Spearman correlation coefficients of −0.511 and 0.427, respectively) and are robust across runs (stds of 1.23 and 1.34%, respectively), while RSS's percent of significant hits is both strongly correlated with the seed set average weighted degree (Spearman 0.945) and much more sensitive to the input seed set (std 12.46%) (**Figure 4**).

A Telomere-Length Maintenance Case Study

In order to study the biological implications of the different normalization methods, we used a telomere length maintenance (TLM) data set from yeast. Specifically, we used a seed set of known TLM genes from Askree et al. (2004) (see Methods and **Supplementary Table S1**). We compiled lists of top-ranking proteins by looking at the top 30 proteins for each of the methods (for RSS, RSS_SD, and RDPN we used $n = 5000$ to increase the resolution of *p*-values produced). We then manually evaluated the relevance of these predicted proteins to telomere length maintenance based on the literature (**Table 2**). We found that the basic propagation produced 4 TLM-related proteins (out of 30), EC produced 5, DADA produced 11,

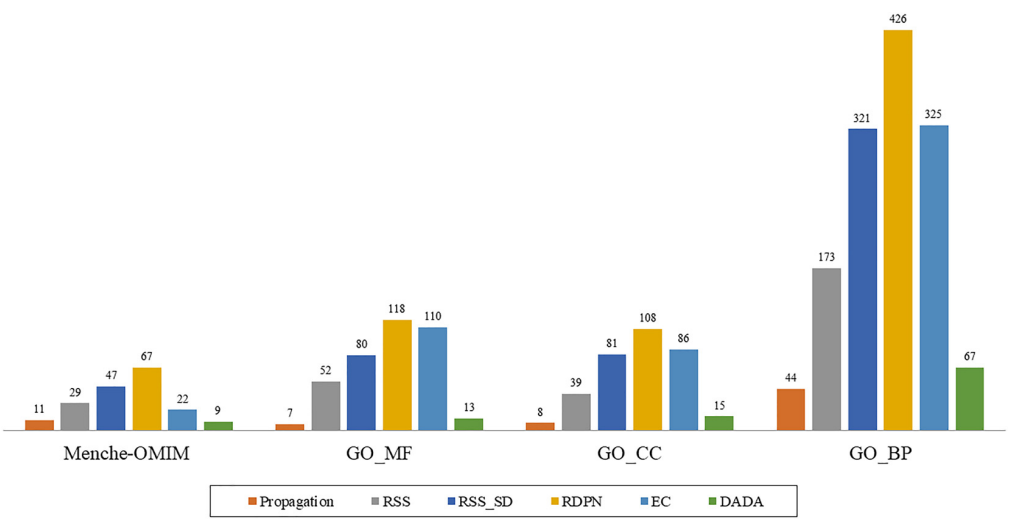


FIGURE 2 | "Best method" counts, based on the AUROC measure, of the six methods across four data sets: Menche-OMIM (173 diseases), GO-MF (358 terms), GO-CC (306 terms), and GO-BP (1237 terms).

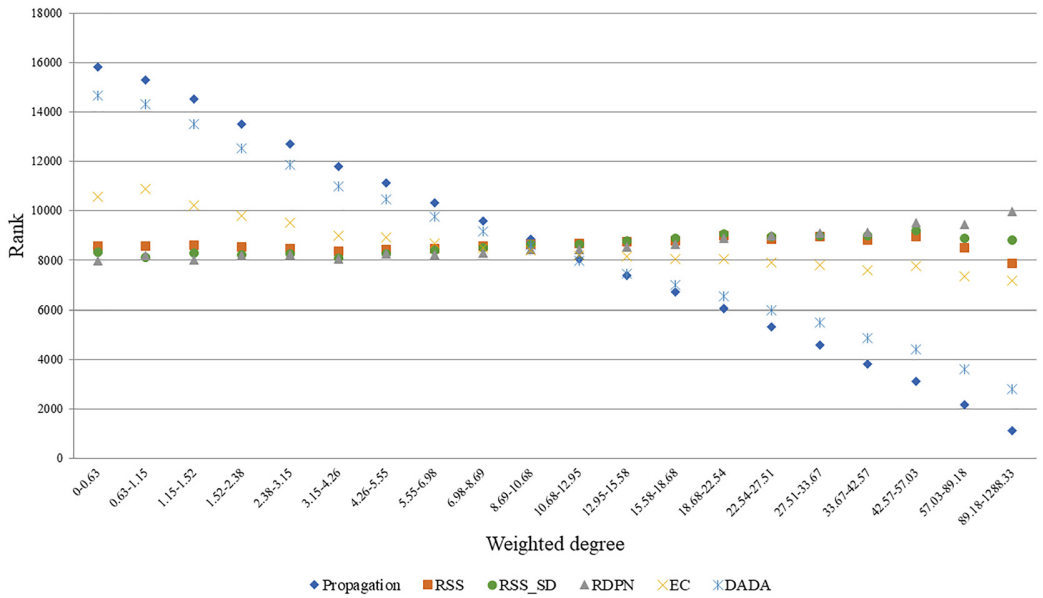


FIGURE 3 | Average rank vs. weighted degree of candidate proteins. Depicted here are ranks based on seed sets from five arbitrary diseases in the Menche-OMIM set (Menche et al., 2015); bins contain approximately equal numbers of proteins. Ranks are derived from the methods' scores the better the score the lower the rank.

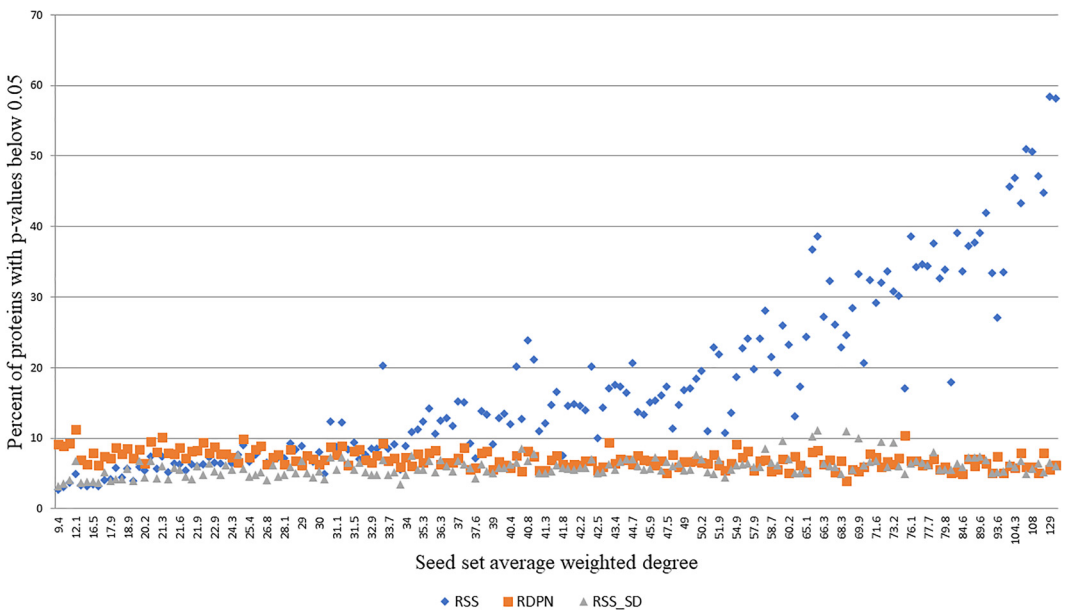


FIGURE 4 | Percent of proteins with *p*-values below 0.05 vs. seed set average weighted degree, using 173 seed sets from the Menche-OMIM data set (Menche et al., 2015).

RSS produced 10, RSS_SD produced 12 and RDPN produced 25. This high specificity (25/30) highlights again the advantage of the newly suggested normalization over previous ones. The newly identified proteins participate in telomere length maintenance as part of large complexes or pathways, such as the VPS pathway, the THO, Mediator and RPD3 complex. The RDPN procedure correctly identified known proteins of these complex previously not characterized. Moreover, out of the 5 proteins not known to be involved in telomere length maintenance, two of them (RNH202 and RNH203) encode subunits of the Rnase H, a nuclease with important roles in genome maintenance, mutated in the human Aicardi-Goutieres syndrome (Crow et al., 2006). Its roles in R-loop repair have suggested possible involvement in telomere biology, although no clear telomere length defect has been detected (Lafuente-Barquero et al., 2017).

TABLE 2 | Top 30 proteins obtained by the different methods in the telomere-length maintenance case study.

	Propagation	EC	DADA	RSS	RSS_SD	RDPN
1	VPS20[15]	LIP2	VPS20[15]	TFG2	SAE2[8,13]	VPS24[1,10]
2	SSB1	RNH203	SRN2[1,10]	SCW10	GBP2[7,14]	SDS3[5]
3	SSA1	RPI1	SSA1	RPB3	TEX1[6]	SRN2[1,10]
4	RPN11	RNH202	SSB1	SUB2	HRB1[4]	MGM1
5	HHT1	PMT5	RNH203	DOA4[12]	THO2[6]	THO2[6]
6	SRN2[1,10]	SRN2[1,10]	RPN11	CPR7	VPS20[15]	RSC8[16]
7	CRM1	RFU1	RNH202	RPO21	CPR7	VPS21[15]
8	HHT2	FLO11	HHT1	GBP2[7,14]	PAF1	VPS20[15]
9	HHF1	SPL2	CRM1	RSC8[16]	SUB2	GAL11[12]
10	HSP82	MVB12	MGM1	DLT1	RAP1[3]	RPO21
11	CDC28	VPS20[15]	HHT2	UBP16	SRN2[1,10]	VPS41[1,10]
12	RNH203	MGM1	HHF1	SUP35	BUD17	MED2[2]
13	RSP5	FMS1	HSP82	VPS24[1,10]	OLA1	GBP2[7,14]
14	RNH202	NTG2	RSP5	RAP1[3]	RIM8	VPS33[1,10]
15	SSB2	SAY1	VPS24[1,10]	HRB1[4]	MTG2	SRB6[2]
16	RPO21	SCW10	RPO21	TEX1[6]	RSC8[16]	MED7[2]
17	HHF2	YKR051W	PEP5	HTB1	RPI1	PEP5
18	DSN1	BSC1	VPS16[1,10]	GAL11[12]	SUP35	VPS8[1,10]
19	MGM1	YBR063C	CDC28	HTA2	RSC3	RXT2[5]
20	CMR1	VPS24[1,10]	SSB2	SCP160	VPS8[1,10]	RNH203
21	VPS24[1,10]	PUT3	THO2[6]	YPK9	DOA4[12]	MED8[2]
22	RVB1	MLH3	HHF2	HHT2	MVB12	VPS4[1,10]
23	RVB2	IBA57	DSN1	NTG2	PEP5	RGR1[16]
24	TOM1	CIA2	VPS33[1,10]	STH1	ALG3	VPS16[1,10]
25	RPC82	MHF1	VPS41[1,10]	HHF1	REB1	DOA4[12]
26	SSC1	ERD2	CMR1	MRX1	SIR2[9,11]	RNH202
27	PEP5	BUD17	SRB4[2]	RGR1[16]	RSC9	CTI6[5]
28	SRB4[2]	CTF8[12]	GAL11[12]	YPR202W	TFG2	HRB1[4]
29	HTA2	RIM8	RGR1[16]	SIR4[12]	YJL070C	RAP1[3]
30	MMS22	VPS38[1,10]	MED8[2]	SRB4	SCW10	TEX1[6]

Proteins in green are related to the TLM mechanism by the following explanations or references: [1]TLM, belongs to the VPS pathway; [2]part of the mediator complex (with SRB2, SRB3, SRB8, SSN2, SSN3, SSN8, GAL11, MED1, NUT1, PGD1, RGR1, and all TLMs); [3]this is the main telomere-length determining protein; [4]paralog of GBP2, the telomere-binding protein; [5]part of RPD3 complex, as DEP1, SAP30, and SIN3 (TLMs); [6]part of the THO/TREX complex (with THP2, HPR1, MFT1 and SOH1, and all TLMs); [7]telomere binding protein; [8]regulator of the MRX complex that processes telomeres; [9]affects telomere chromatin, although not telomere length; [10]Dieckmann et al. (2016); [11]Ellahi et al. (2015); [12]Gatbonton et al. (2006); [13]Hardy et al. (2014); [14]Konkel et al. (1995); [15]Shachar et al. (2008); [16]Ungar et al. (2009).

CONCLUSION

In summary, we have devised a new method (RDPN) for normalizing propagation results that accounts for the degrees of the involved proteins and produces robust *p*-value estimations. The method was shown to outperform previous ones across diverse disease-related and function-related data sets. Importantly, we have shown that the *p*-values it assigns do not depend on the degree of the protein being scored, hence this method is less prone to literature biases and more likely to discover new associations. Moreover, we have shown that its assigned *p*-values are robust to the average degree of the seed set, allowing significance assessment across different data sets. Finally, in testing the biological implications of the method's predictions, we found that it greatly outperforms previous normalizations and leads to new biological insights.

Considering all evaluated parameters, it seems that three of the tested methods outshine the others: RDPN, which generates robust *p*-values and displays the best performance, RSS_SD which also generates robust *p*-values but doesn't perform as well, and EC which is easy to implement and has good performance although its nominal scores are harder to interpret.

We note that there are many variants in the literature of the basic network propagation methodology, such as random walk with restart and diffusion kernel (Cowen et al., 2017). Our normalization method is readily applicable to all these variants and can be used to eliminate potential degree biases and assign statistical significance values.

METHODS

Normalization Methods

Normalization With Random Seed Sets (RSS)

This method uses propagation scores from n random seed sets (we use $n = 100$ unless stated otherwise) to normalize the real

propagation scores, as suggested by Mazza et al. (2016). In detail, each protein v has a "real" propagation score X^v_{real} the score it got by propagating from the real seed set; and n random scores X^v_i ($0 \leq i \leq n-1$) derived by propagating from n random seed sets (each with the same number of proteins as the real seed set). For every protein v only the instances in which it was not part of the random seed set are considered, and its p-value is the fraction of random instances in which its score exceeded its real propagation score, i.e.:

$$p^v = \frac{|\{i|(X^v_i \geq X^v_{real} \text{ and } v \text{ was not part of the } i\text{'th random seed set})\}| + 1}{|\{i|(v \text{ was not part of the } i\text{'th random seed set})\}| + 1}$$

Normalization With Eigenvector Centrality (EC)

The EC scores are computed as follows:

$$p^v = \frac{X^v_{\alpha=0.8}}{X^v_{\alpha=1}}$$

where $X^v_{\alpha=0.8}$ is the propagation score of protein v when propagating from the seed set with $\alpha = 0.8$, and $X^v_{\alpha=1}$ is its propagation score when propagating from the same seed set with $\alpha = 1$ (i.e., disregarding the seed set in the computation).

DADA

The DADA ranks, as described in Erten et al. (2011), are computed as follows: first EC scores are computed as:

$$EC^v = log\left(\frac{X^v_{\alpha=0.7}}{X^v_{\alpha=1}}\right)$$

for all the proteins in the network where $X^v_{\alpha=0.7}$ is the propagation score of protein v when propagating from the seed set with $\alpha = 0.7$, and $X^v_{\alpha=1}$ is its propagation score when propagating from the same seed set with $\alpha = 1$. Then each protein gets a rank R^i_{EC} which is its position in a descending order of EC scores, and also a rank R^v_{prop} which is its position in a descending order of the regular propagation scores $X^v_{\alpha=0.7}$. Finally, if the average weighted degree of the seed set exceeds the network average weighted degree, all proteins final ranks are set to R^v_{prop}. Otherwise, they are set to R^v_{EC}.

Normalization With Random Similar Degree Distributed Seed Sets (RSS_SD)

Following Erten et al. (2011), we first construct seed sets $S(i)$ ($0 \leq i \leq n-1$, we use $n = 100$) that have a degree distribution that is similar to the original seed set S by applying this procedure: We assign each $v \in V$ to a bucket $B(u)$ such that $u \in S$ and $|W(v) - W(u)|$ is minimized (ties are broken randomly).

In case there are two or more seed proteins with an equal weighted degree, there is a possibility that one of their buckets will remain empty. If that happens, we reassign all network proteins (we repeat this step if necessary).

We generate $S(i)$ by choosing a protein from each bucket uniformly at random.

We then propagate from these seed sets, as well as from the original seed set, and proceed to compute p-values as in the RSS method.

Data Sets

Menche-OMIM Data Set

Menche et al. (2015) compiled a list of 299 diseases defined by the Medical Subject Headings (MeSH) that have at least 20 associated genes from either the Online Mendelian Inheritance in Man (OMIM) data set or the genome-wide association study (GWAS) data set (or both). We empirically found that all methods perform better when using only the genes from OMIM, so only the 173 diseases out of that list that have at least 20 and up to 1000 associated genes from OMIM in the HIPPIE network were used for evaluation.

GO Data Set

We used geneSCF (Subhash and Kanduri, 2016) to get a list of all GO terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2017) (in all three sub-ontologies) with their corresponding genes. We focused the evaluation on terms that included between 20 and 1000 genes (1237 GO Biological Process (BP) terms, 306 GO Cellular Component (CC) terms and 358 GO Molecular Function (MF) terms).

TLM Data Set

A genome wide-screen study by Askree et al. (2004) found 173 *S. cerevisiae* genes that affect telomere length. We used 163 of them that are found in the ANAT *S. cerevisiae* network as the seed set (**Supplementary Table S1**).

PPI Networks

For the performance evaluation section we used the HIPPIE network which has 17335 proteins and 330028 (non self-loops) interactions in its main connected component (Alanis-Lobato et al., 2017) (version 18-Jul-2017).

For the TLM case study we used the ANAT *Saccharomyces cerevisiae* network which has 5527 proteins and 75678 (non self-loops) interactions in its main connected component (Almozlino et al., 2017).

Area Under ROC Curve (AUROC) Measure

For each group of disease-related or function-related genes, we randomly split it to five equally sized parts. In each cross-validation iteration we hid one of the parts, used the other four as a seed set, and tested the success of the method in predicting the hidden proteins (serving as positive samples) using the AUROC measure. We then averaged the performance across the five iterations. To compute the AUROC scores, we picked negative samples with similar weighted degrees as the positive samples. This was implemented as follows: for each positive protein with a weighted degree w, we chose the smallest integer r such that there are at least 100 proteins in the network (excluding the

seed set, the positive samples and the already chosen negative samples) with weighted degree in the range *[w−r, w+r]*. We then randomly picked a protein from this group to be used as a negative sample.

AUTHOR CONTRIBUTIONS

HB and RS conceived the RDPN method and designed the computational framework. HB implemented the framework and produced the results. All authors interpreted the results and contributed to the manuscript.

FUNDING

RS was supported by the Israel Science Foundation (Grants No. 715/18 and 757/12). MK was supported by the Israel Science Foundation and the Israel Cancer Research Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00004/full#supplementary-material

REFERENCES

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi: 10.1093/nar/gkw985

Almozlino, Y., Atias, N., Silverbush, D., and Sharan, R. (2017). ANAT 2.0: reconstructing functional protein subnetworks. *BMC Bioinformatics* 18:495. doi: 10.1186/s12859-017-1932-1

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Askree, S. H., Yehuda, T., Smolikov, S., Gurevich, R., Hawk, J., Coker, C., et al. (2004). A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc. Natl. Acad. Sci. U. S. A.* 101, 8658–8663. doi: 10.1073/pnas.0401263101

Biran, H., Almozlino, T., Kupiec, M., and Sharan, R. (2018). WebPropagate: a web-server for network propagation. *J. Mol. Biol.* 430, 2231–2236. doi: 10.1016/j.jmb.2018.02.025

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* 30, 107–117. doi: 10.1016/S0169-7552(98)00110-X

Bryan, K., and Leise, T. (2006). The $25,000,000,000 eigenvector: the linear algebra behind google. *SIAM Rev.* 48, 569–581. doi: 10.1137/050623280

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18:551. doi: 10.1038/nrg.2017.38

Crow, Y. J., Leitch, A., Hayward, B. E., Garner, A., Parmar, R., Griffith, E., et al. (2006). Mutations in genes encoding ribonuclease H2 subunits cause aicardi-goutières syndrome and mimic congenital viral brain infection. *Nat. Genet.* 38, 910–916. doi: 10.1038/ng1842

Dieckmann, A. K., Babin, V., Harari, Y., Eils, R., König, R., Luke, B., et al. (2016). Role of the ESCRT complexes in telomere biology. *mBio* 7, e01793–e01816. doi: 10.1128/mBio.01793-16

Ellahi, A., Thurtle, D. M., and Rine, J. (2015). The chromatin and transcriptional landscape of native *Saccharomyces cerevisiae* telomeres and subtelomeric domains. *Genetics* 200, 505–521. doi: 10.1534/genetics.115.175711

Erten, S., Bebek, G., Ewing, R. M., and Koyutürk, M. (2011). DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* 4:19. doi: 10.1186/1756-0381-4-19

Gatbonton, T., Imbesi, M., Nelson, M., Akey, J. M., Ruderfer, D. M., Kruglyak, L., et al. (2006). Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.* 2:e35. doi: 10.1371/journal.pgen.0020035

Hardy, J., Churikov, D., Géli, V., and Simon, M. N. (2014). Sgs1 and Sae2 promote telomere replication by limiting accumulation of ssDNA. *Nat. Commun.* 5:5004. doi: 10.1038/ncomms6004

Konkel, L. M., Enomoto, S., Chamberlain, E. M., McCune-Zierath, P., Iyadurai, S. J., and Berman, J. (1995). A class of single-stranded telomeric DNA-binding proteins required for Rap1p localization in yeast nuclei. *Proc. Natl. Acad. Sci. U. S. A.* 92, 5558–5562. doi: 10.1073/pnas.92.12.5558

Lafuente-Barquero, J., Luke-Glaser, S., Graf, M., Silva, S., Gómez-González, B., Lockhart, A., et al. (2017). The Smc5/6 complex regulates the yeast Mph1 helicase at RNA-DNA hybrid-mediated DNA damage. *PLoS Genet.* 13:e1007136. doi: 10.1371/journal.pgen.1007136

Mazza, A., Klockmeier, K., Wanker, E., and Sharan, R. (2016). An integer programming framework for inferring disease complexes from network data. *Bioinforma. Oxf. Engl.* 32, i271–i277. doi: 10.1093/bioinformatics/btw263

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J., and Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences. arXiv:cond-mat/0312028 [Preprint].

Shachar, R., Ungar, L., Kupiec, M., Ruppin, E., and Sharan, R. (2008). A systems-level approach to mapping the telomere length maintenance gene circuitry. *Mol. Syst. Biol.* 4:172. doi: 10.1038/msb.2008.13

Subhash, S., and Kanduri, C. (2016). GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics* 17:365. doi: 10.1186/s12859-016-1250-z

The Gene Ontology Consortium. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Ungar, L., Yosef, N., Sela, Y., Sharan, R., Ruppin, E., and Kupiec, M. (2009). A genome-wide screen for essential yeast genes that affect telomere length maintenance. *Nucleic Acids Res.* 37, 3840–3849. doi: 10.1093/nar/gkp259

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6:e1000641. doi: 10.1371/journal.pcbi.1000641

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.

The reviewer MK declared a past collaboration with one of the authors RS.

Copyright © 2019 Biran, Kupiec and Sharan. 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) and the copyright owner(s) 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 Genetics

REVIEW
published: 01 May 2019
doi: 10.3389/fgene.2019.00381

To Embed or Not: Network Embedding as a Paradigm in Computational Biology

Walter Nelson[1,2], Marinka Zitnik[3], Bo Wang[3,4,5], Jure Leskovec[3,6], Anna Goldenberg[1,5,7] and Roded Sharan[8]**

[1] Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada, [2] Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada, [3] Department of Computer Science, Stanford University, Stanford, CA, United States, [4] Peter Munk Cardiac Center, University Health Network, Toronto, ON, Canada, [5] Vector Institute, Toronto, ON, Canada, [6] Chan Zuckerberg Biohub, San Francisco, CA, United States, [7] Department of Computer Science, University of Toronto, Toronto, ON, Canada, [8] School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Current technology is producing high throughput biomedical data at an ever-growing rate. A common approach to interpreting such data is through network-based analyses. Since biological networks are notoriously complex and hard to decipher, a growing body of work applies graph embedding techniques to simplify, visualize, and facilitate the analysis of the resulting networks. In this review, we survey traditional and new approaches for graph embedding and compare their application to fundamental problems in network biology with using the networks directly. We consider a broad variety of applications including protein network alignment, community detection, and protein function prediction. We find that in all of these domains both types of approaches are of value and their performance depends on the evaluation measures being used and the goal of the project. In particular, network embedding methods outshine direct methods according to some of those measures and are, thus, an essential tool in bioinformatics research.

Keywords: network biology, network embedding, network alignment, community detection, protein function prediction

OPEN ACCESS

Edited by:
Marco Pellegrini, Italian National Research Council (CNR), Italy

Reviewed by:
Noel Malod-Dognin, Barcelona Supercomputing Center, Spain
Gregorio Alanis-Lobato, Francis Crick Institute, United Kingdom

****Correspondence:***
Anna Goldenberg
anna.goldenberg@utoronto.ca; anna.goldenberg@gmail.com
Roded Sharan
roded@tau.ac.il

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *05 February 2019*
Accepted: *09 April 2019*
Published: *01 May 2019*

Citation:
Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A and Sharan R (2019) To Embed or Not: Network Embedding as a Paradigm in Computational Biology. Front. Genet. 10:381. doi: 10.3389/fgene.2019.00381

INTRODUCTION

Network biology is a powerful paradigm for representing, interpreting and visualizing biological data (Barabási and Oltvai, 2004). One of the standard approaches to computing on networks is to transform such data into vectorial data, aka *network embedding*, to facilitate similarity search, clustering and visualization (Hamilton et al., 2017b; Cai et al., 2018).

In a network embedding problem, one is given a network and an induced similarity (or distance) function between its nodes; the goal is to find a low dimensional representation of the network nodes in some metric space so that the given similarity (or distance) function is preserved as much as possible. For example, if the input network is unweighted and the distance between nodes is defined to be the graph geodesic distance, then a possible goal could be to find an embedding into Euclidean space that minimizes the sum of squared differences between graph distances and the corresponding Euclidean distances (Tenenbaum, 2000).

The classical approach to network embedding employs matrix factorization and is based on the fact that if the desired similarity matrix is positive semi-definite then it can be decomposed into the

product of a real matrix and its transpose. Thus, if one represents each node by a row of that matrix then the given similarity is completely captured by the dot-product between the corresponding vector representations. Similarly, if one is given distances between nodes that satisfy the triangle inequality then double centering the distance matrix gives a positive semi-definite matrix whose decomposition yields vector representations that respect the given distances. This approach is precisely the multidimensional scaling procedure (Cox and Cox, 2000).

Embedding approaches have several potential advantages. Algorithms making use of embeddings are frequently faster than their counterparts which operate on the original networks. Additionally, the learned embeddings are often applicable for downstream analysis, either by direct interpretation of the embedding space or through the application of machine learning techniques which are designed for vectorial data. Beyond its computational advantages, network embedding is natural to use in biological problems that concern physical entities (such as proteins) that function in 3D space. In such scenarios, Euclidean representations may capture many of the functional properties of those entities. Finally, by working in lower dimensional space, the results are more likely to be robust to the noise inherently present in the networks. Indeed, recent network denoising approaches employed embedding for this purpose (Wang et al., 2018).

In this review, we describe several current approaches for graph embedding including spectral-based, diffusion-based and deep-learning-based methods. We provide comparisons applying representative embedding approaches to fundamental problems in network biology with using the networks directly in three distinct tasks: protein network alignment, protein module detection, and protein function prediction (**Figure 1**). We further review network embedding methods and their application to network denoising and pharmacogenomics. We conclude that network embedding methods are an essential component in the bioinformatics tool box.

METHODOLOGY

Methods for network embedding aim to optimize the difference between the node similarities/distances in the original network space and their similarities/distances under the embedding, which is typically constrained to have a low dimension. In the following, we describe various methods for embedding a given network in Euclidean space. For a graph G with n nodes, a weighted adjacency matrix W and a diagonal degree matrix D, we define its Laplacian matrix as $L = D\text{-}W$.

Graph drawing algorithms are perhaps the best-known embedding techniques, commonly used to visualize a graph in 2D space. Initially proposed in (Eades, 1984) as an extension of (Tutte, 1963), and further developed in (Fruchterman and Reingold, 1991), the spring-embedder model is a particularly elegant example: one can imagine that connected pairs of nodes are attached to springs which bring them closer together, while all nodes repel each other so as not to be placed too closely together. Other classes of graph drawing algorithms, including multi-level and dimensionality reduction-based techniques, are described in detail in a recent review (Gibson et al., 2013). Spatial analysis of functional enrichment (Baryshnikova, 2018) is one recent application of force-directed graph drawing algorithm, designed for the annotation and visualization of large, complex biological networks.

One of the fundamental methods to decompose a matrix is spectral decomposition, i.e., decomposing the matrix into its eigenvectors and eigenvalues. Given a network, the principal eigenvectors Q of its Laplacian matrix capture membership of nodes in implicit network clusters, commonly used for embedding (Belkin and Niyogi, 2003). The matrix Q is obtained by optimizing $min_{Q\in R^{n\times C}} Trace\left(Q^{T}L^{+}Q\right)$, s.t. $Q^{T}Q = I$, where $L^{+} = \text{I-D}^{-1/2}WD^{-1/2}$ is a normalized Laplacian and C is the number of clusters. However, this spectral embedding reflects the global structure in the network without taking into consideration more fine-grained local structures and is therefore sensitive to noise. Wang et al. (2017a) recently introduced the Vicus matrix as a local-neighborhood version of the Laplacian matrix. Each cell of the Vicus matrix represents the probability of node j having the same label as node i if we did a random walk around the local neighborhood of node i. Encoding local neighborhoods in this fashion does not only preserve the geometric properties of the original Laplacian matrix but also reduces the noise and improves the quality of the embedding. Wang et al. showed that for a variety of tasks, including network clustering of single-cell RNA-seq data, cluster stability, identification of rare cell populations, and ranking of genes associated with cancer subtypes, Vicus-based spectral methods outperformed Laplacian-based spectral methods on a wide variety of biological tasks.

Diffusion-based approaches focus on embedding nodes into low-dimensional vector spaces by first using random walks to construct a network neighborhood of every node in the network, and then optimizing an objective function with network neighborhoods as input (Perozzi et al., 2014a; Tang et al., 2015; Grover and Leskovec, 2016). The objective function is carefully designed to preserve both the local and global network structures. For example, a popular method, Mashup, complements traditional random walks, which yield only diffusion states, with a dimensionality reduction step that is aimed at reducing the noise in these diffusion computations. To this end, Mashup approximates each diffusion state s_i with a multinomial logistic model based on a latent vector representation of nodes that uses far fewer dimensions than the original, n-dimensional state. Specifically, if the latent vector representation for node i is denoted by x_i, Mashup also constructs a contextual vector w_i that has the same dimensionality as x_i and captures the topology of the subnetwork around node i. To this end, Mashup computes the probability assigned to node j in the diffusion state of node i as $\widehat{s_{ij}} = \frac{exp(x_i^T w_j)}{\sum_k exp(x_i^T w_k)}$, so that these computed diffusion states align with the original diffusion states. Mashup constructs an optimization framework to minimize the KL-divergence of these two diffusion states and applies standard gradient descent methods to solve for the latent representations.

Another widely used network embedding algorithm that uses random walks is node2vec (Grover and Leskovec, 2016).

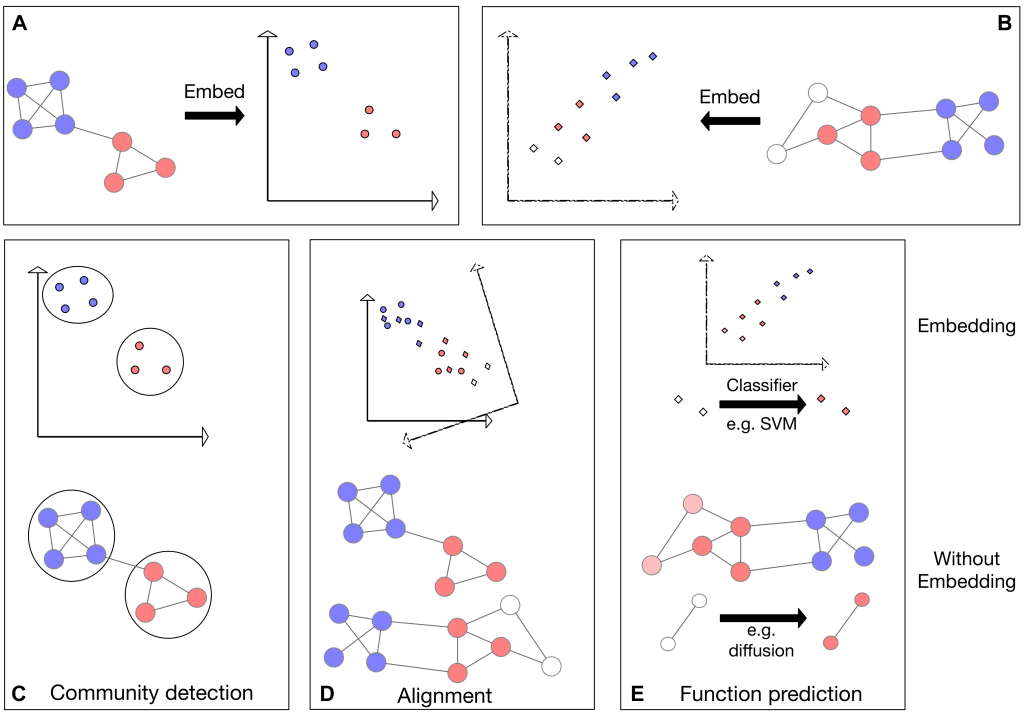


FIGURE 1 | Schematic representing three applications applied to networks directly as well as applied to the network embeddings. Colors represent some node features in the network; for example, protein families. **(A,B)** Visualization of the embedding process for two networks in 2D space. **(C)** Visualization of community detection in embedded space (*top*) and directly on the network (*bottom*). **(D)** *Top*: visualization of network alignment in embedded space. In this example, the network embedding in panel **(B)** is rotated, translated and reflected to find an optimal alignment with the embedding in panel **(A)**. *Bottom*: visualization of direct alignment of two networks: vertical proximity represents the found alignment. **(E)** Visualization of function prediction in embedded space. The previously unlabeled (white) nodes (*bottom*) or their embeddings (*top*) are labeled (colored).

Node2vec learns node embeddings so that a node's embedding can predict nearby (neighborhood) nodes. Technically, the network neighborhood $N(u)$ is a set of nodes that appear in an appropriately biased, short random walk from node u (Grover and Leskovec, 2016). The goal of the algorithm is to find an embedding $f(u)$ such that the conditional probability of observing u's network neighbors $N(u)$ is maximized. This conditional probability is modeled using a softmax function, leading to the following log likelihood: $\sum_{u} \sum_{v \in N(u)} log \frac{exp(f(u) \cdot f(v))}{\sum_{w} exp(f(w) \cdot f(u))}$, across all nodes u in the network. Once embeddings are learned, one can use them for any downstream prediction task, including node classification, link prediction, and clustering. A similar network embedding algorithm is DeepWalk (Perozzi et al., 2014b). DeepWalk has been originally proposed to embed nodes in a social network setting, taking ideas from the linguistics literature (Perozzi et al., 2014b). In DeepWalk, the embeddings are learned based on truncated random walks which can be intuitively thought of as putting words (nodes) into sentences (sequences of nodes visited by a random walk). In the biological context, DeepWalk has been used to associate miRNAs with diseases (Li et al., 2017), predict drug target associations (Zong et al., 2017), and predict protein function (Kulmanov et al., 2017).

With the advent of deep learning methods, several deep learning approaches were proposed to embed networks. An important class of deep learning methods for network embedding are graph neural networks that generalize the notion of convolutions typically applied to image datasets to operations that can operate on arbitrary graphs (Defferrard et al., 2016; Kipf and Welling, 2016; Gilmer et al., 2017; Hamilton et al., 2017a). One can see graph neural networks as an embedding methodology that distills high-dimensional information about each node's neighborhood into a dense vector embedding without requiring manual feature engineering (Defferrard et al., 2016; Kipf and Welling, 2016; Gilmer et al., 2017; Hamilton et al., 2017a). A graph neural network has two main components. First, the encoder, maps a node u to a low-dimensional embedding $f(u)$, based on u's local neighborhood structure, its position in the graph, and/or its attributes. Next, the decoder takes the embeddings and extracts user-specified predictions from these embeddings. In contrast to embedding approaches that use random walks (reviewed above), graph neural networks support end-to-end learning. One can jointly optimize all trainable parameters and propagate gradients of the objective function through the encoder as well as the decoder. End-to-end learning can lead to substantial improvements in performance (Defferrard et al., 2016; Zitnik et al., 2018).

There has been significant recent interest in graph embeddings in non-Euclidean spaces. In particular, hyperbolic spaces have attracted much attention due to successful natural language

processing models which use them for embedding words (Chamberlain et al., 2017). Muscoloni et al. (2017) describe a general algorithm termed "coalescent embedding" for embedding vertices in hyperbolic spaces. The algorithm proceeds by pre-weighting the network and applying a non-linear dimension reduction technique, followed by computing and adjusting the angular positions of the Euclidean embeddings and radial positioning according to node degree. More generally, networks and their respective embeddings can be interpreted geometrically, as described in recent reviews (Barthélemy, 2011; Papadopoulos et al., 2015; Moyano, 2017). These geometric models have been used successfully in applications to biological networks, particularly protein–protein interaction (PPI) networks (Serrano et al., 2012; Alanis-Lobato et al., 2016, 2018).

APPLICATIONS

Network Alignment

A basic operation in biological research is to transfer knowledge across species. Indeed, sequence alignment has been the power horse of computational biology for almost five decades now. With the availability of physical interaction data, it was suggested to generalize alignment concepts to the network level (Kelley et al., 2003; Sharan and Ideker, 2006). There are several types of network alignment problems, here we focus on global network alignment where given the networks of two species (typically, PPI networks) one wishes to identify a 1–1 correspondence between the proteins of the two species under which the networks are most similar (**Figure 1D**).

A leading approach to this problem is the IsoRank algorithm (Singh et al., 2008) which is based on Google's PageRank method, essentially measuring the correspondence, or similarity, between two proteins from different species based on the similarities of their neighboring nodes in the two corresponding networks. Thus, if we denote by R_{ij} the similarity between proteins i and j (from two different species), and we let $N(i)$ denote the (open) neighborhood of protein i in its network, then:

$$R_{ij} = \frac{1}{|N(i)||N(j)|} \sum_{u \in N(i), v \in N(j)} R_{uv}$$

These recursive equations give rise to an eigenvalue problem and their solution is used as an input to a maximum matching algorithm to compute the eventual correspondence.

Another, more recent approach is MAGNA (Saraph and Milenkoviæ, 2014) and its successor MAGNA++ (Vijayan et al., 2015). MAGNA uses a genetic algorithm to find the optimal alignment, where individuals are viewed as permutations of the nodes. Crossover relies on the notion of adjacency, where a pair of permutations is adjacent if they differ only by a single swap of two nodes; the crossover of two permutations is then the midpoint of the shortest path between the two permutations in the graph constructed from these adjacencies. Selection can be based on any metric, such as EC. MAGNA++ augments this approach by including cross-species node similarity information. An extensive review of methods for biological network alignment can be found in (Guzzi and Milenkovic, 2018) that mentions over thirty different approaches. Comparative network analysis methods are further reviewed in (Emmert-Streib et al., 2016).

A recent work by Fan et al. (2017) uses an embedding-based approach, MuNK, to compare networks across species by assessing similarity via embedded network topologies. The idea is to project the nodes of the two networks into the same Euclidean space in a way that preserves their intra-species network similarity and inter-species sequence similarity. For each species separately, a kernel similarity function is defined, and the corresponding embedding is computed by matrix decomposition. To tie the projections together, Fan et al. (2017) assume a given set of known matches, regarded as landmarks, between the two networks. A similar embedding approach that does not require a known subset of correspondences was suggested in (Heimann et al., 2018).

As a test case for network embedding, we evaluated the two algorithms, IsoRank and MuNK, using metrics of alignment quality. A common and simple metric is the edge correctness (EC), defined as the percentage of edges conserved under the mapping f (Kuchaiev et al., 2009; Clark and Kalita, 2014):

$$EC = \frac{|f(E_A) \cap E_B|}{|E_A|} \times 100\%$$

Note that the EC metric is asymmetric, and the order of the networks is traditionally chosen to maximize EC, i.e., A is chosen to be the smaller of the two networks. Beyond topological similarity, one can use different biological annotations, such as the Gene Ontology (GO) functional annotation, to compute biologically relevant measures of alignment quality such as GO functional consistency (Aladag and Erten, 2013), defined as the proportion of aligned pairs with more than k GO terms in common.

Similar to the use of landmarks in MuNK, IsoRank can incorporate additional similarity information in its computation of the score matrix, so the landmark pairs are provided as a binary information matrix to the IsoRank algorithm. In our experiments, we produce two outputs for method comparison: cross-species pairwise similarity scores and the node-to-node mappings. Thus, in addition to the two measures described above that use the node-to-node mappings, we also evaluated IsoRank and MuNK using AUPR as a measure of enrichment of GO functional consistency with respect to the cross-species pairwise similarity scores. When comparing MuNK to the more recent MAGNA++, MAGNA++ performs very well according to EC (as it optimizes EC directly), but it does not output node scores so we could not directly compare MuNK to MAGNA++ according to AUPR and other metrics. Per the author recommendation, the regularization parameter for the Laplacian in MuNK was fixed at $\lambda = 0.05$. Damping can be used in the PageRank step of the IsoRank algorithm, and therefore we performed a grid search with step size 0.05 over possible convexity parameters $\alpha \in (0, 1)$, optimizing for EC score. As input data, we use the PPI networks for two species of yeast, *S. cerevisiae* and

S. pombe, extracted from the BioGRID interaction database (Oughtred et al., 2018).

IsoRank performs better on the measures directly related to the node mapping (**Table 1**). This may be due to the fact that the cross-species similarity coefficients in IsoRank directly incorporate local neighborhood (i.e., topological) information, a fact that the IsoRank greedy algorithm is designed to take advantage of. The MuNK scores predict functional correctness better than the scores produced by IsoRank, suggesting that MuNK's learned embedded space is biologically meaningful potentially even beyond alignment. In comparing network alignment methods (Guzzi and Milenkovic, 2018) also found that methods that do very well according to the topological quality measures are not very good as far as functional quality is concerned. The interpretability of the embedding space is one of the primary benefits of embedding techniques over standard approaches in the case of network alignment. For example, the embedding space learned by MuNK captures biological information beyond pairwise node alignment, specifically, cross-species synthetic lethal interactions (Fan et al., 2017).

Community Detection

One of the natural uses of a network is the identification of clusters, or modules of similar nodes, a task known as community detection (Fortunato, 2010). Community detection methods (**Figure 1C**) have great uses in biology from protein module identification to disease subnetwork discovery (Ghiassian et al., 2015; Menche et al., 2015). Among the most popular community detection methods on networks are random walk-based approaches including Louvain (Blondel et al., 2008), Infomap (Rosvall and Bergstrom, 2011), Label propagation (Raghavan et al., 2007), and Walktrap (Pons and Latapy, 2005), that came up as best performers in a review comparing these and other approaches (Yang et al., 2016). Originally developed for community detection in social networks, these methods are frequently used in biology (Barabási et al., 2011), for example to identify cancer drivers (Cantini et al., 2015).

Network embedding for the purpose of community detection was covered in a recent review (Hamilton et al., 2017b). The authors hypothesized that due to vector-like embedding representation of a network, there is a wider range of clustering and community detection methods that can be applied to embedded networks as compared to graphs directly. The authors further introduced an encoder-decoder framework that unifies many of the recently popularized approaches, including DeepWalk (Perozzi et al., 2014a) and node2vec (Grover and Leskovec, 2016). A geometric approach, not covered in the review, suggests a scalable embedding of networks in a hyperbolic circle and show that the popular random walk-based community detection methods (Louvain, Infomap, Label propagation, and Walktrap) can be significantly boosted when applied to hyperbolic distances (Muscoloni et al., 2017).

TABLE 1 | Comparative analysis of direct vs. embedding methods across a range of problems in network biology.

	IsoRank ($\alpha = 0.5$)	**MuNK** ($\lambda = 0.05$)
A. Network alignment		
EC	**39.0%**	21.9%
GOC		
$K = 20$	**63.4%**	57.6%
$K = 50$	**20.7%**	17.9%
$K = 100$	**1.2%**	1.0%
GOC (AUPR)	0.721	**0.746**
Runtime	26 min 40 s (incl. grid search)	1 min 52 s (incl. alignment)
	densityCut ($K = 4$, $\alpha = 0.9$)	**Vicus** ($K = 10$, $\sigma = 0.5$)
B. Community detection		
Buettner ($C = 2$, $C_t = 11$)	0.256	**0.316**
Kolodziejczyk ($C = 5$, $C_t = 4$)	0.325	**0.552**
Pollen ($C = 13$, $C_t = 11$)	**0.931**	0.928
Usoskin ($C = 9$, $C_t = 4$)	0.373	**0.591**
Avg. Runtime	1 min 15 s (incl. parameter grid search)	<5 s
STRING v9.1 *Homo sapiens*. Included with the Mashup distribution.		
	GeneMANIA	**Mashup**
C. Function prediction		
AUPR		
MF	0.327	**0.372**
BP	0.213	**0.222**
CC	**0.514**	0.487
Avg. Runtime	3 min 57 s	14 min 56 s (incl. recommended SVM tuning procedure)

Running times were obtained on a 64-bit machine with Intel Core i5-8400 CPU @ 2.80 GHz × 6 with 16 GB RAM running Ubuntu 18.04. Bold refers to the most successful result, according to the referenced metric.

We compared two community detection methods, an embedding-based and a graph-based, on the problem of single-cell RNA-seq (scRNA-seq) analysis. scRNA-seq data has recently emerged as a powerful tool to decipher the heterogeneity of cell populations. This is an important and growing area of network applications where community detection methods are used to perform clustering on the constructed cell-to-cell networks (Wang et al., 2018). Given a gene expression matrix, Gaussian kernel is usually adopted to construct a pairwise similarity network in which nodes represent cells and edge weights depict the similarity between cells.

The first method is Vicus, a generalization of spectral clustering, which we combined with *k*-means clustering in the embedded space. For the network-based approach, we used densityCut, a random walk-based community detection method, which approximates clusters using the density of local neighborhoods. The densityCut method approximates the true network using a *k*-nearest neighbor graph, and selects the number of clusters using an automated procedure. Therefore, this number of clusters was used as input to the *k*-means step of the Vicus evaluation. We used four scRNA-seq datasets, all from *Mus musculus* (Pollen et al., 2014; Buettner et al., 2015; Kolodziejczyk et al., 2015; Usoskin et al., 2015) but which vary according to tissue of origin (neural, blood and stem cells) and have known ground truth labels. We evaluated performance using normalized mutual information (NMI). Vicus outperformed densityCut on three of the four datasets (**Table 1**).

Function Prediction

Another fundamental problem in network biology is the inference of protein function from the known functions of its network neighbors (Sharan et al., 2007). The earliest approach to this problem, neighborhood counting (Schwikowski et al., 2000), predicted a protein to be involved in a certain function if a sufficient number of its direct (or up to some specified distance) neighbors had this property. Current state of the art methods are based on similar guilt-by-association principles (**Figure 1E**). For example, Cao et al. (2013) define a distance metric between proteins that is based on network diffusion, thus capturing similarities that are based on multiple paths in the network.

These single-network methods were generalized in several ways (Cho et al., 2016) integrate information across multiple networks and use a low rank approximation of the network diffusion based similarities to reduce potential noise. The integration challenge is also tackled by (Gligorijevic et al., 2018) who learn a compact node representation using deep autoencoders. In Fan et al. (2017), the cross-species embedding is utilized to infer protein function. Zitnik and Leskovec (2017) suggest a network embedding approach for predicting tissue-specific protein function, which encourages proteins to share features not only with their network neighbors but also with proteins that are active in similar tissues.

Two recent methods were compared on the task of protein function prediction using multiple interaction networks. GeneMANIA performs label diffusion, while Mashup finds an embedding for each of the proteins, allowing one to use traditional classification techniques such as support vector machines (SVMs). The area under the precision-recall curve (AUPR) was used as an evaluation metric. Overall, Mashup performed better with respect to molecular function and biological process annotations, while GeneMANIA performed better on the cellular compartment annotation (**Table 1**).

Network Denoising

The application of network biology techniques to experimental data depends on the accuracy and completeness of the network of interest. The challenge of noisy interaction measurements plagues many different types of biological networks, such as Hi-C interaction networks (Rao et al., 2014), cell–cell interaction networks (Wang et al., 2017b), and PPI networks (Saito et al., 2002; Przulj et al., 2004; Chua and Wong, 2008; Higham et al., 2008; Kuchaiev et al., 2009; You et al., 2010; Marras et al., 2011; Alanis-Lobato et al., 2013; Cannistraci et al., 2013; Newman, 2018a,b). Such noise adversely impacts the performance of downstream analysis, calling for methods for network denoising.

The most common approach to denoise any given network is to perform diffusions on the network to exploit high-order structures that can potentially improve the qualities of the direct links between nodes. Diffusion maps (Coifman et al., 2005) employ high-order random walks and then use spectral decomposition to construct an affinity measure. A tensor-based dynamical model (Wang et al., 2012) aims to search high-order paths between pairs of objects through their common nearest neighbors. A low-rank constraint has been employed to help denoise the network manifold (Wang and Tu, 2013). Diffusion-state distance (DSD) (Cao et al., 2013) was utilized to denoise PPI networks and improve the signal-to-noise ratio for better prediction of protein functions. To tackle the problem of transitive edges in networks in a computationally efficient way (Feizi et al., 2013) proposed a simple closed-form solution, called Network Deconvolution (ND), to infer direct links.

An alternative direction of network denoising takes embedding-based approaches. For instance, Mashup (Cho et al., 2016) aims to learn compact low-dimensional vector representation of proteins that best explains their wiring patterns for the input protein–protein association networks by applying a matrix factorization method on the diffused network. The embeddings of the nodes (proteins) reflect the relational structures of the original network, therefore facilitating the downstream applications by feeding the embeddings to a support vector machine.

A recent study (Wang et al., 2018) performed an in-depth comparison between these network denoising methods in three different experimental settings: PPI function predictions, HiC network module detection, and species identification. The study highlighted the advantages of embedding-based methods such as Mashup (Cho et al., 2016) when the network contains distinct cluster structures and the noise level is small. However, it also showed that when the cluster structures are corrupted by high

noise, existing methods usually fail to uncover the underlying network structure.

Pharmacogenomics

Modern pharmaceutical research faces challenges with decreasing productivity in drug development and a persistent gap between therapeutic needs and available treatments (Hodos et al., 2016; Moffat et al., 2017). Network approaches have emerged as a promising direction to address these challenges and improve our understanding of the therapeutic and side effects of drugs (Hopkins, 2008; Berger and Iyengar, 2009). We review three practically important problems within the realm of pharmacogenomics that have been tackled with network embedding methods: drug-target prediction, drug–drug interaction prediction and prediction problems involving small molecules.

Drugs influence biological systems by binding to target proteins and affecting their downstream activity (Imming et al., 2006). Network approaches formulate drug–target interaction prediction as a link prediction task on a graph of drugs/chemicals and the proteins which they interact with (Yildirim et al., 2007; Yamanishi et al., 2010; Perlman et al., 2011; Chen et al., 2012; Cheng et al., 2012; Gönen, 2012; Isik et al., 2015; Zitnik and Zupan, 2016; Luo et al., 2017; Wen et al., 2017; Lee and Nam, 2018). Given such a graph (Crichton et al., 2018) use various node embedding methods, including node2vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014b), and LINE (Tang et al., 2015), to embed nodes into a compact vector space in a manner that preserves local network structure. As a result, drugs with many shared target proteins obtain similar embeddings, and vice-versa, proteins targeted by similar drugs obtain similar embeddings. These embeddings are thus well-suited for predicting drug–target interactions by calculating the similarity between embeddings representing the drug and the protein, or by using embeddings as inputs to a machine learning method (Crichton et al., 2018). Alternatively, predictions can be made in an end-to-end fashion, where a neural network learns node embeddings and predicts interactions directly from the graph (Wang and Zeng, 2013; Gao et al., 2018; Wan et al., 2018).

TABLE 2 | A summary of network embedding tools and their applications.

Name of the tool	Availability	What was it applied to
Denoising		
Network enhancement	Matlab code http://snap.stanford.edu/ne/	Hi-C interaction networks combining gene interaction networks across tissues
Single-cell representation learning	Binary https://github.com/SuntreeLi/SCRL	Single-cell RNA-seq data
Geometric denoising	http://kuchaev.com/Denoising/	PPI networks
Network alignment		
MuNK	Python code and all Anaconda-reproducible experiments https://github.com/lrgr/munk	Cross-species functional PPIs (yeast, mouse, human)
Community detection		
Minimum curvilinearity embedding II	https://sites.google.com/site/carlovittoriocannistraci/5-datasets-and-matlab-code/minimum-curvilinearity-ii-april-2012	(i) Cerebrospinal fluid proteomics – neuropathic pain
		(ii) Transcription factor expressions – tissue prediction
Vicus		Single-cell RNA-seq:
		(i) Pollen – neural and stem cells
		(ii) Usoskin – mouse neurons, sensory subtypes
		(iii) Buettner – embryonic stem cells
		(iv) Kolodziejczyk – pluripotent cells
Coalescent embedding	https://github.com/biomedical-cybernetics/coalescent_embedding	Non-biological
Function prediction		
Mashup	http://cb.csail.mit.edu/cb/mashup/	Protein function prediction, gene ontology reconstruction, and genetic interaction prediction
OhmNet	http://snap.stanford.edu/ohmnet/	Tissue-specific gene function prediction
Disease gene discovery	http://snap.stanford.edu/pathways/	Disease pathway detection
Pharmacogenomics		
Molecular fingerprints	https://github.com/HIPS/neural-fingerprint	Prediction of molecular properties, including drug efficacy, solubility, and photovoltaic efficiency
Decagon	http://snap.stanford.edu/decagon/	(i) Polypharmacy side-effect prediction
		(ii) Drug–drug interaction prediction
Graph convolutional policy network	https://github.com/bowenliu16/rl_graph_generation	Molecular graph generation
Residual LSTM Embeddings	https://github.com/deepchem/deepchem	(i) Drug side-effect prediction
		(ii) Drug toxicity prediction

Detecting drug–drug interactions, in which the activity of one drug changes, favorably or unfavorably, if taken with another drug, is an important challenge with significant implications for patient mortality and morbidity (Chan and Giaccia, 2011; Guthrie et al., 2015; Han et al., 2017). Ma et al. (2018) model each drug as a node in a multi-view drug association graph, where edges between drugs in different views encode different types of similarity between drugs. The approach uses graph convolutional networks (Kipf and Welling, 2016) to embed the multi-view graph and attentive mechanisms (Veličković et al., 2018) to fuse information from multiple views and to make learning more interpretable. By such embedding, the approach learns a similarity score between any two drugs and uses the scores to predict drug–drug interactions. While such an approach can be useful to describe drug interactions at the cellular level (Sridhar et al., 2016; Ryu et al., 2018), it cannot predict the safety or side effects of drug combinations. To identify the side effects of drug combinations and provide guidance on the development of new drug therapies (Zitnik et al., 2018) developed an embedding approach that constructs a multi-modal graph of PPIs, drug–protein interactions, and drug–drug interactions, where each drug–drug interaction is labeled by a different edge type signifying the type of the side effect. The approach takes the multi-modal graph and uses graph neural networks as an embedding methodology to distill information about each node's network neighborhood into an embedding vector without any hand-engineering. The final approach is an end-to-end method for predicting side effects of drug combinations that considers all types of side effects at once. The approach learns embeddings of side effects that are indicative of polypharmacy in patients.

Chemical prediction problems represent another class of practically important graph problems (Ralaivola et al., 2005; Altae-Tran et al., 2017; Gilmer et al., 2017; Gómez-Bombarelli et al., 2018). One key distinction between these problems and standard network prediction tasks discussed above is that chemical prediction problems are graph-level classification problems where individual data examples are graphs (rather than nodes) representing small molecules. Typical prediction tasks aim to predict various molecular properties such as drug efficacy or solubility (Coley et al., 2017; Jin et al., 2017), predict which drugs bind to which target proteins (Morris et al., 2018), and identify sites at which a particular candidate drug binds to a target protein (Feinberg et al., 2018). The input to a predictor is a small molecule, which is commonly represented as a graph in which nodes and edges represent atoms and bonds between atoms, respectively. One difficulty with such inputs is that molecular graphs can be of arbitrary size and shape (Niepert et al., 2016; Xu et al., 2017). However, currently, most machine learning pipelines can only handle inputs of a fixed size. For this reason, state-of-the-art systems use embedding techniques to embed molecular graphs into fixed-dimensional embeddings and then use the learned representations as inputs to a fully connected deep neural network or other standard machine learning methods (Duvenaud et al., 2015; Kearnes et al., 2016). The proposed graph convolution models do not yet consistently outperform traditional structural-based fingerprints, however, their flexibility and potential for further optimization and development have led to models that provide significant boosts in the predictive power over older fingerprints.

CONCLUSION

We have reviewed several classes of approaches for network embedding, including spectral-based methods, random-walk based approaches and deep neural network techniques. We have demonstrated the utility of these approaches in a broad set of applications, ranging from network alignment to community detection, protein function prediction, and network denoising. We have also discussed recent embedding approaches in pharmacogenomics. We were interested in seeing whether the field of network embedding indeed enhances the types of questions that can be answered using graph-based approaches and our conclusion is that there is value in both graph-based and graph-embedding-based methods in a variety of applications.

In our experiments we found that depending on the task at hand and metric used, sometimes graph-based methods outperformed network embedding tools. This was the case with, for example, IsoRank beating MuNK with respect to edge conservation in network alignment, whereas MuNK outperformed IsoRank according to the area under the precision recall curve with respect to node mapping. In community detection experiments, our results were reversed, where the embedding method outperformed the graph-based method 3 out of 4 times. In fact, there is no single metric according to which one type of method is consistently better than the other. Even in compute time, where embedding methods outperform graph-based methods most of the time, on the function prediction task graph-based GeneMANIA outperforms the embedding method Mashup. This implies that the choice of graph-based versus embedding-based method will depend on many factors, not just the task at hand, but also the aspect or evaluation measure of highest importance to the user.

The network embedding principles create new opportunities to model large network datasets and move beyond standard prediction tasks of node classification, link prediction, and node clustering. For example, given a partially observed network of interactions between drugs, diseases, and proteins, one might be interested in posing a logical query: "What proteins are likely to be associated with diseases that have both symptoms X and Y?" Such a query requires reasoning about all possible proteins that might be associated with at least two diseases, which, in turn, clinically manifest through symptoms X and Y. Valid answers to such queries correspond to subgraphs. Since edges in the network might be missing because of biotechnological limits and natural variation, naively answering the queries requires enumeration over all possible combinations of diseases (Hamilton et al., 2018) developed a network embedding approach that answers such complex logical queries and achieves a time complexity linear in the size of a query, compared to the exponential complexity required by a naive enumeration-based approach. The approach embeds nodes into a low-dimensional space and

represents logical operators as learned geometric operations in this embedding space. They demonstrated the utility of the approach in a study involving a biomedical network of drugs, diseases, proteins, side effects, and protein functions with millions of edges.

We summarize network embedding tools that are used in the biomedical field in **Table 2**. We expect the importance of these tools to grow with the magnitude and complexity of biomedical data that are being generated.

AUTHOR CONTRIBUTIONS

WN did the performance comparisons. All authors participated in writing the manuscript.

FUNDING

AG and RS were supported by a TAU-UOT cooperation grant.

REFERENCES

Aladag, A. E., and Erten, C. (2013). SPINAL: scalable protein interaction network alignment. *Bioinformatics* 29, 917–924. doi: 10.1093/bioinformatics/btt071

Alanis-Lobato, G., Cannistraci, C. V., and Ravasi, T. (2013). Exploitation of genetic interaction network topology for the prediction of epistatic behavior. *Genomics* 102, 202–208. doi: 10.1016/j.ygeno.2013.07.010

Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. (2018). The latent geometry of the human protein interaction network. *Bioinformatics* 34, 2826–2834. doi: 10.1093/bioinformatics/bty206

Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. A. (2016). Manifold learning and maximum likelihood estimation for hyperbolic network embedding. *Appl. Netw. Sci.* 1:10. doi: 10.1007/s41109-016-0013-0

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3, 283–293. doi: 10.1021/acscentsci.6b00367

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Barthélemy, M. (2011). Spatial networks. *Phys. Rep.* 499, 1–101.

Baryshnikova, A. (2018). Spatial analysis of functional enrichment (SAFE) in large biological networks. *Methods Mol. Biol.* 1819, 249–268. doi: 10.1007/978-1-4939-8618-7_12

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Berger, S. I., and Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics* 25, 2466–2472. doi: 10.1093/bioinformatics/btp465

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008:10008.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160. doi: 10.1038/nbt.3102

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* 30, 1616–1637. doi: 10.1109/tkde.2018.2807452

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013). Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* 29, i199–i209. doi: 10.1093/bioinformatics/btt208

Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep.* 5: 17386.

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., et al. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS One* 8:e76339. doi: 10.1371/journal.pone.0076339

Chamberlain, B. P., Clough, J., and Deisenroth, M. P. (2017). Neural embeddings of graphs in hyperbolic space. arXiv:1705.10359 [Preprint]. doi: 10.1371/journal.pone.0076339

Chan, D. A., and Giaccia, A. J. (2011). Harnessing synthetic lethal interactions in anticancer drug discovery. *Nat. Rev. Drug Discov.* 10, 351–364. doi: 10.1038/nrd3374

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8, 1970–1978.

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8:e1002503. doi: 10.1371/journal.pcbi.1002503

Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* 3, 540.e5–548.e5.

Chua, H. N., and Wong, L. (2008). Increasing the reliability of protein interactomes. *Drug Discov. Today* 13, 652–658. doi: 10.1016/j.drudis.2008.05.004

Clark, C., and Kalita, J. (2014). A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* 30, 2351–2359. doi: 10.1093/bioinformatics/btu307

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., et al. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7426–7431. doi: 10.1073/pnas.0500334102

Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S., and Jensen, K. F. (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* 57, 1757–1772. doi: 10.1021/acs.jcim.6b00601

Cox, T. F., and Cox, M. A. A. (2000). *Multidimensional Scaling*, 2nd Edn. Boca Raton, FL: CRC Press.

Crichton, G., Guo, Y., Pyysalo, S., and Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics* 19:176. doi: 10.1186/s12859-018-2163-9

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)* (Barcelona: NIPS), 3844–3852.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints," in *Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015)* (Barcelona: NIPS), 2224–2232.

Eades, P. (1984). A heuristic for graph drawing. *Congr. Numer.* 42, 149–160.

Emmert-Streib, F., Dehmer, M., and Shi, Y. (2016). Fifty years of graph matching, network alignment and network comparison. *Inf. Sci.* 34, 180–197. doi: 10.1016/j.ins.2016.01.074

Fan, J., Cannistra, A., Fried, I., Lim, T., Schaffner, T., Crovella, M., et al. (2017). A multi-species functional embedding integrating sequence and network structure. *bioRxiv* [Preprint].

Feinberg, E. N., Sur, D., Husic, B. E., Mai, D., Li, Y., Yang, J., et al. (2018). Spatial graph convolutions for drug discovery. arXiv:1803.04465 [Preprint].

Feizi, S., Marbach, D., Médard, M., and Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31, 726–733. doi: 10.1038/nbt.2635

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Fruchterman, T. M. J., and Reingold, E. M. (1991). *Graph Drawing by Force-Directed Placement*. Hoboken, NJ: Wiley.

Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., and Zhang, P. (2018). "Interpretable drug target prediction using deep neural representation," in *Proceedings of the Conference: Twenty-Seventh International Joint Conference on Artificial Intelligence* (New York, NY: IJCAI), 3371–3377.

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Gibson, H., Faith, J., and Vickers, P. (2013). A survey of two-dimensional graph layout techniques for information visualisation. *Inf. Visual.* 12, 324–357. doi: 10.1177/1473871612455749

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. arXiv:1704.01212v2 [Preprint].

Gligorijevic, V., Barot, M., and Bonneau, R. (2018). deepNF: deep network fusion for protein function prediction. *Bioinformatics* 34, 3873–3881. doi: 10.1093/bioinformatics/bty440

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Gönen, M. (2012). Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 28, 2304–2310. doi: 10.1093/bioinformatics/bts360

Grover, A., and Leskovec, J. (2016). "node2vec: scalable feature learning for networks," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, Vol. 22 (New York, NY: ACM), 855–864.

Guthrie, B., Makubate, B., Hernandez-Santiago, V., and Dreischulte, T. (2015). The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010. *BMC Med.* 13:74. doi: 10.1186/s12916-015-0322-7

Guzzi, P. H., and Milenkovic, T. (2018). Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Brief. Bioinform.* 19, 472–481.

Hamilton, W. L., Bajaj, P., Zitnik, M., Jurafsky, D., and Leskovec, J. (2018). Embedding logical queries on knowledge graphs. arXiv:1806.01445v3 [Preprint].

Hamilton, W. L., Ying, R., and Leskovec, J. (2017a). Inductive representation learning on large graphs. *arxiv*.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017b). Representation learning on graphs: methods and applications. *arxiv*.

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., and Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.* 35, 463–474. doi: 10.1038/nbt.3834

Heimann, M., Shen, H., Safavi, T., and Koutra, D. (2018). "REGAL: representation learning-based graph alignment," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management – CIKM '18* (New York, NY: ACM), doi: 10.1145/3269206.3271788

Higham, D. J., Rašajski, M., and Pržulj, N. (2008). Fitting a geometric graph to a protein–protein interaction network. *Bioinformatics* 24, 1093–1099. doi: 10.1093/bioinformatics/btn079

Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., and Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 8, 186–210. doi: 10.1002/wsbm.1337

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi: 10.1038/nchembio.118

Imming, P., Sinning, C., and Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* 5, 821–834. doi: 10.1038/nrd2132

Isik, Z., Baldow, C., Cannistraci, C. V., and Schroeder, M. (2015). Drug target prioritization by perturbed gene expression and network information. *Sci. Rep.* 5:17417.

Jin, W., Coley, C., Barzilay, R., and Jaakkola, T. (2017). "Predicting organic reaction outcomes with weisfeiler-lehman network," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (New York, NY: Curran Associates, Inc.), 2607–2616.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* 30, 595–608. doi: 10.1007/s10822-016-9938-8

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., et al. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11394–11399. doi: 10.1073/pnas.1534710100

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv:1609.02907v4 [Preprint].

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C. H., Ilicic, T., Henriksson, J., Natarajan, K. N., et al. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17, 471–485. doi: 10.1016/j.stem.2015.09.011

Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009). Geometric de-noising of protein-protein interaction networks. *PLoS Comput. Biol.* 5:e1000454. doi: 10.1371/journal.pcbi.1000454

Kulmanov, M., Khan, M. A., Hoehndorf, R., and Wren, J. (2017). DeepGO: predicting protein functions from sequence using a deep ontology-aware classifier. *Bioinformatics* 34, 660–668. doi: 10.1093/bioinformatics/btx624

Lee, I., and Nam, H. (2018). Identification of drug-target interaction by a random walk with restart method on an interactome network. *BMC Bioinformatics* 19:208. doi: 10.1186/s12859-018-2199-x

Li, G., Luo, J., Xiao, Q., Liang, C., Ding, P., and Cao, B. (2017). Predicting MicroRNA-disease associations using network topological similarity based on deepwalk. *IEEE Access* 5, 24032–24039. doi: 10.1109/access.2017.2766758

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8:573.

Ma, T., Xiao, C., Zhou, J., and Wang, F. (2018). Drug similarity integration through attentive multi-view graph auto-encoders. arXiv:1804.10850v1 [Preprint].

Marras, E., Travaglione, A., and Capobianco, E. (2011). Manifold learning in protein interactomes. *J. Comput. Biol.* 18, 81–96. doi: 10.1089/cmb.2009.0258

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. (2017). Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.* 16, 531–543. doi: 10.1038/nrd.2017.111

Morris, P., DaSilva, Y., Clark, E., Hahn, W. E., and Barenholtz, E. (2018). "Convolutional neural networks for predicting molecular binding affinity to HIV-1 proteins," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* BCB '18 (New York, NY: ACM), 220–225.

Moyano, L. G. (2017). Learning network representations. *Eur. Phys. J. Spec. Top.* 226, 499–518.

Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G., and Cannistraci, C. V. (2017). Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat. Commun.* 8:1615.

Newman, M. E. J. (2018a). Estimating network structure from unreliable measurements. *Phys. Rev. E* 98:062321.

Newman, M. E. J. (2018b). Network structure from rich but noisy data. *Nat. Phys.* 14, 542–545. doi: 10.1038/s41567-018-0076-1

Niepert, M., Ahmed, M., and Kutzkov, K. (2016). "Learning convolutional neural networks for graphs," in *Proceedings of the International Conference on Machine Learning* (jmlr.org), Bejing, 2014–2023.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., et al. (2018). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541.

Papadopoulos, F., Aldecoa, R., and Krioukov, D. (2015). Network geometry inference using common neighbors. *Phys. Rev. E* 92:022807.

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.* 18, 133–145. doi: 10.1089/cmb.2010.0213

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014a). "Deepwalk: online learning of social representations," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)* (New York, NY: ACM), 701–710.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014b). "DeepWalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '14 (New York, NY: ACM), 701–710.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular

heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Pons, P., and Latapy, M. (2005). "Computing communities in large networks using random walks," in *Computer and Information Sciences – ISCIS 2005 Lecture Notes in Computer Science*, eds P. Yolum, T. Güngör, F. Gürgen, and C. Özturan (Berlin: Springer), 284–293. doi: 10.1007/11569596_31

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 3508–3515. doi: 10.1093/bioinformatics/bth436

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76:036106.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Netw.* 18, 1093–1110. doi: 10.1016/j.neunet.2005.07.009

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Rosvall, M., and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One* 6:e18209. doi: 10.1371/journal.pone.0018209

Ryu, J. Y., Kim, H. U., and Lee, S. Y. (2018). Deep learning improves prediction of drug–drug and drug-food interactions. *Proc. Natl. Acad. Sci. U.S.A.* 115, E4304–E4311.

Saito, R., Suzuki, H., and Hayashizaki, Y. (2002). Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res.* 30, 1163–1168. doi: 10.1093/nar/30.5.1163

Saraph, V., and Milenkoviæ, T. (2014). MAGNA: maximizing accuracy in global network alignment. *Bioinformatics* 30, 2931–2940. doi: 10.1093/bioinformatics/btu409

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261. doi: 10.1038/82360

Serrano, M. Á, Boguñá, M., and Sagués, F. (2012). Uncovering the hidden geometry behind metabolic networks. *Mol. Biosyst.* 8, 843–850.

Sharan, R., and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* 24, 427–433. doi: 10.1038/nbt1196

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3:88.

Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12763–12768. doi: 10.1073/pnas.0806627105

Sridhar, D., Fakhraei, S., and Getoor, L. (2016). A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics* 32, 3175–3182. doi: 10.1093/bioinformatics/btw342

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web* WWW '15 (Geneva: International World Wide Web Conferences Steering Committee), 1067–1077.

Tenenbaum, J. B. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Tutte, W. T. (1963). How to draw a graph. *Proc. Lond. Math. Soc.* 3, 743–767.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153. doi: 10.1038/nn.3881

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. arXiv:1710.10903v3 [Preprint].

Vijayan, V., Saraph, V., and Milenkoviæ, T. (2015). MAGNA++: maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics* 31, 2409–2411. doi: 10.1093/bioinformatics/btv161

Wan, F., Hong, L., Xiao, A., Jiang, T., and Zeng, J. (2018). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *bioRxiv* [Preprint].

Wang, B., Huang, L., Zhu, Y., Kundaje, A., Batzoglou, S., and Goldenberg, A. (2017a). Vicus: exploiting local structures to improve network-based analysis of biological data. *PLoS Comput. Biol.* 13:e1005621. doi: 10.1371/journal.pcbi.1005621

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017b). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14, 414–416. doi: 10.1038/nmeth.4207

Wang, B., Jiang, J., Wang, W., Zhou, Z., and Tu, Z. (2012). "Unsupervised metric fusion by cross diffusion," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC: IEEE Computer Society), 2997–3004.

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S., et al. (2018). Network enhancement as a general method to denoise weighted biological networks. *Nat. Commun.* 9:3108.

Wang, B., and Tu, Z. (2013). "Sparse subspace denoising for image manifolds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 468–475.

Wang, Y., and Zeng, J. (2013). Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* 29, i126–i134. doi: 10.1093/bioinformatics/btt234

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., et al. (2017). Deep-learning-based drug-target interaction prediction. *J. Proteome Res.* 16, 1401–1409. doi: 10.1021/acs.jproteome.6b00618

Xu, Z., Wang, S., Zhu, F., and Huang, J. (2017). "Seq2Seq fingerprint: an unsupervised deep molecular embedding for drug discovery," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* ACM-BCB '17 (New York, NY: ACM), 285–294.

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254. doi: 10.1093/bioinformatics/btq176

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* 6:30750.

Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.* 25:1119.

You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* 26, 2744–2751. doi: 10.1093/bioinformatics/btq510

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, 457–466.

Zitnik, M., and Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* 33, i190–i198. doi: 10.1093/bioinformatics/btx252

Zitnik, M., and Zupan, B. (2016). "Collective pairwise classification for multi-way analysis of disease and drug data," in *Proceedings of the Pacific Symposium on Biocomputing* (Bethesda MD: NIH Public Access), 81.

Zong, N., Kim, H., Ngo, V., and Harismendy, O. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics* 33, 2337–2344. doi: 10.1093/bioinformatics/btx160

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.

Copyright © 2019 Nelson, Zitnik, Wang, Leskovec, Goldenberg and Sharan. 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) and the copyright owner(s) 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 Genetics

REVIEW
published: 14 June 2019
doi: 10.3389/fgene.2019.00549

Comprehensive Review of Models and Methods for Inferences in Bio-Chemical Reaction Networks

Pavel Loskot[1]*, *Komlan Atitey*[1] and *Lyudmila Mihaylova*[2]

[1] College of Engineering, Swansea University, Swansea, United Kingdom, [2] Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, United Kingdom

The key processes in biological and chemical systems are described by networks of chemical reactions. From molecular biology to biotechnology applications, computational models of reaction networks are used extensively to elucidate their non-linear dynamics. The model dynamics are crucially dependent on the parameter values which are often estimated from observations. Over the past decade, the interest in parameter and state estimation in models of (bio-) chemical reaction networks (BRNs) grew considerably. The related inference problems are also encountered in many other tasks including model calibration, discrimination, identifiability, and checking, and optimum experiment design, sensitivity analysis, and bifurcation analysis. The aim of this review paper is to examine the developments in literature to understand what BRN models are commonly used, and for what inference tasks and inference methods. The initial collection of about 700 documents concerning estimation problems in BRNs excluding books and textbooks in computational biology and chemistry were screened to select over 270 research papers and 20 graduate research theses. The paper selection was facilitated by text mining scripts to automate the search for relevant keywords and terms. The outcomes are presented in tables revealing the levels of interest in different inference tasks and methods for given models in the literature as well as the research trends are uncovered. Our findings indicate that many combinations of models, tasks and methods are still relatively unexplored, and there are many new research opportunities to explore combinations that have not been considered—perhaps for good reasons. The most common models of BRNs in literature involve differential equations, Markov processes, mass action kinetics, and state space representations whereas the most common tasks are the parameter inference and model identification. The most common methods in literature are Bayesian analysis, Monte Carlo sampling strategies, and model fitting to data using evolutionary algorithms. The new research problems which cannot be directly deduced from the text mining data are also discussed.

Keywords: automation, Bayesian analysis, biochemical reaction network, estimation, inference, modeling, survey, text mining

OPEN ACCESS

Edited by:
Marco Pellegrini, Institute of Computer Science and Telematics (IIT), Italy

Reviewed by:
Adriano Velasque Werhli, Fundação Universidade Federal do Rio Grande, Brazil
Jiri Vohradsky, Institute of Microbiology (ASCR), Czechia

***Correspondence:**
Pavel Loskot
p.loskot@swan.ac.uk

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *08 February 2019*
Accepted: *24 May 2019*
Published: *14 June 2019*

Citation:
Loskot P, Atitey K and Mihaylova L (2019) Comprehensive Review of Models and Methods for Inferences in Bio-Chemical Reaction Networks. Front. Genet. 10:549. doi: 10.3389/fgene.2019.00549

1. INTRODUCTION

Biological systems are presently subject to extensive research efforts to ultimately control the underlying biological processes. The challenge is the level of complexity of these systems with intricate dependencies on the internal and external conditions. Biological systems are inherently non-linear, dynamic as well as stochastic. Their responses to input perturbations are often

difficult to predict as they may respond differently to the same inputs. Moreover, biological phenomena must be considered at different spatio-temporal scales, from single molecules to gene-scale reaction networks.

Many biological systems can be conveniently represented as biological circuits (Zamora-Sillero et al., 2011), or as networks of biochemical reactions (Ashyraliyev et al., 2009). Common examples of biological systems which can be described as BRNs are: metabolic networks, signal transduction networks, gene regulatory networks (GRNs), and more generally, the networks of biochemical pathways. Moreover, BRNs share similar characteristics with evolutionary and prey-predatory networks in population biology, and disease spreading networks in epidemiology. Synthetic bio-reactors and other types of chemical reactors used in industrial production are other examples of BRNs (Ali et al., 2015).

Qualitative as well as quantitative observations of biological systems are necessary to elucidate their functional and structural properties. Despite the advent of high throughput experiments, the biological phenomena are often only partially observed. Since the internal system state cannot be fully nor directly observed, it must be inferred from the measurements. Such inferences are possible due to the dependency of observations on the internal states and parameter values (Fröhlich et al., 2017). Single molecule techniques are promising for advancing the cell biology as they enable more focused observations, however, their resolution and dimensionality is still limited.

The observations in experiments are often distorted and noisy, and involve some form of averaging. Extended models can be assumed for the measurements involving distortion (Ruttor and Opper, 2009). The measurement noise may not be additive nor Gaussian, and its variance may be dependent on the values of other parameters. The parameter values may differ for *in vitro* and *in vivo* experiments (Famili et al., 2005). In systems comprising chemical reactions, the parameters of interest are usually initial and instantaneous concentrations, reaction rates and possibly other kinetic constants including the diffusion and drift coefficients. The molecular concentrations can be usually measured directly whereas the other parameters must be inferred from measurements (Fröhlich et al., 2017). The parameter inferences as well as measurements can be performed sequentially (online) or in batches (off-line) (Arnold et al., 2014).

In BRNs, the number of chemical species is usually much smaller than the number of chemical reactions. In some cases, it may be useful to estimate the number of reactions between consecutive measurements (Reinker et al., 2006). The structural identifiability of a chemical reaction system is affected by which reactions are occurring.

The observations at possibly non-equidistant time instances represent longitudinal data which can be used to create or validate mathematical models. The rate of discrete time observations is important (Fearnhead et al., 2014), since more frequent observations can be costly, and affect the observed biological processes. Processing the large volumes of data is also computationally demanding. The observations and their processing can be merged to create so-called observers in order to replace the high-cost sensors in chemical reactors (Rapaport and Dochain, 2005). Observers can be classified as explanatory or predictive to describe the existing or future data, respectively (Ali et al., 2015). Observers can process discretized and delayed measurements, and yield the interval measurements of quantities with a variable observation gain (Vargas et al., 2014). The average state observers of large-scale systems are defined in Sadamoto et al. (2017).

The dynamics of biological processes can be elucidated from their mathematical models. The importance of modeling in biology is discussed in Chevaliera and Samadb (2011), and general modeling strategies are described in Banga and Canto (2008). The research problems in biology dictate what physical and chemical processes must be included in the models. It is usually more efficient to only collect the observations which are necessary to formulate and test a biological hypothesis than to perform extensive, time consuming and expensive laboratory experiments. Such a strategy is referred to as a forward modeling (Reinker et al., 2006). On the other hand, finding the parameter values to reproduce the observations can be enhanced by the experiment design, and it is known as a reverse modeling (Hagen et al., 2013). The differences between forward and reverse modeling strategies are explained in Ashyraliyev et al. (2009).

The models of biological systems are dependent on the *in vivo* or *in vitro* experiments considered. BRNs can be modeled as deterministic input-output non-linear transformations which can be sometimes locally linearized at a given time scale and resolution. The models can be modified using additional transformations to facilitate their analysis. Apart from deterministic models, there are also stochastic, event-driven and probabilistic models of BRNs. When the number of species is large, the stochastic models converge to deterministic models (Rempala, 2012). The same model used multiple times can represent a biological population (Woodcock et al., 2011).

Biological models need to be unbiased in order to avoid systematic errors. Since they are usually evaluated many times, they need to be computationally fast, and at the right

Abbreviations: ABC, approximate Bayesian computation, artificial bee colony; ABM, agent based model; AR, alternating regression; CCA, canonical correlation analysis; CDIS, conditional density importance sampling; CGA, continuous genetic algorithm; CLE, chemical Langevin equation; CME, chemical master equation; CRO, chemical reaction optimization; CS, compressive sensing; CTMC, continuous time Markov chain; CTMP, continuous time Markov process; DE, differential evolution; DLR, deep learning; EKF, extended Kalman filter; EM, expectation-maximization; EP, expectation propagation; FA, firefly algorithm; FDM, finite differences method; GLR, generalized linear regression; GLR, generalized linear regression; GP, genetic programming; HDL, hardware description language; KF, Kalman filter; LFM, linear fractional model; LNA, linear noise approximation; LS, least squares; MAP, maximum a posterior; MC, Monte Carlo; MCEM, MC expectation-maximization; MCMC, MC Markov Chain; MES, maximum entropy sampling; ML, maximum likelihood; MLR, machine learning; MM, method of moments; MMSE, minimum mean square error; NLP, non-linear programming, natural language processing; NLR, narrative literature review; NLSQ, non-linear least squares; ODE, ordinary differential equation; PDF, portable document format, probability density function; PMC, population Monte Carlo; PSO, particle swarm optimization; QE, quasi-equilibrium; QSS, quasi-steady state; RDME, reaction-diffusion master equation; RRE, reaction rate equation; SA, simulated annealing; SMC, sequential Monte Carlo; SMCMC, sequential Markov chain Monte Carlo; SS, scatter search; SSE, sum of squared errors, system size expansion; SLR, systematic literature review; TLR, transfer learning; UKF, unscented Kalman filter.

level of coarse grain description. For instance, microscopic stochastic models may be computationally expensive whereas a deterministic macroscopic description, such as population-average modeling may not be sufficiently accurate due to a low level of resolution.

Development of large-scale kinetic systems is one of the key tasks in contemporary computational biology (Penas et al., 2017). The corresponding models can be multidimensional and have 100's or even 1000's of parameters, and constraints while the initial conditions are not known. The models can be hierarchical or nested, and have parts interconnected by multiple feedback loops (Rodriguez-Fernandez et al., 2013). The parameter estimation for large-scale reaction networks is considered in Remlia et al. (2017).

The model analysis can yield the transient responses of a biological system, and to obtain the behavior at steady state or in equilibrium (Atitey et al., 2018a). It may be also useful to explore complex multi-dimensional parameter spaces. The viable parameter values of many models of biological systems form only a small fraction of the overall parameter space (Atitey et al., 2019), so identifying this sub-volume by simple sampling is rather inefficient (Zamora-Sillero et al., 2011). The model analysis is further complicated by the size of the state space, the number of unknown parameters, the analytical intractability, and various numerical problems. Evaluation of the observation errors can both facilitate as well as validate the model analysis (Bouraoui et al., 2015).

The majority of analytical and numerical methods can be used universally for models with different structures. The efficiency of model analysis can be considered in the statistical or computational sense. In the statistical sense, the analysis needs to be robust against the uncertainty in model structure and the parameter values estimated from noisy and limited observations. The computational efficiency can be achieved by developing the algorithms which are prone to massively parallel implementations (Nobile et al., 2012).

In this review paper, we are primarily concerned with the parameter inference in biological and chemical systems described by various models of BRNs. We use the terms inference and estimation interchangeably. In the literature, the parameter inference is also referred to as an inverse problem (Engl et al., 2009), point estimation, model calibration and model identification. The key objective of the parameter inference is to minimize a suitably defined estimation error while suppressing the effects of measurement errors (Sadamoto et al., 2017). More recently, machine learning methods are becoming popular as an alternative to learn not only the model parameters, but also to learn the model features from the labeled and unlabeled observations (Sun et al., 2012; Schnoerr et al., 2017).

The parameter inference is affected by many factors. For instance, different models experience a different degree of structural identifiability. Provided that different parameter values or different inputs generate the same dynamic response, such as the statistics of synthesized molecules, the model parameters cannot be identified, or can only be partially identified. In some cases, the structural identifiability can be overcome by changing the modeling strategy (Yenkie et al., 2016). The structural identifiability is a necessary but not sufficient condition for the overall model identifiability (Gábor et al., 2017). A relationship between the identifiability and observability is discussed in Baker et al. (2011). The practical identifiability (also known as a posterior identifiability) assesses whether there is enough data to suppress the measurement noises. It may be beneficial to test the identifiability of the parameters of interest prior to attempting their inference. For instance, the parameters may not be identifiable at a given time scale, or the data may not have sufficient dimensionality (variability) or volume. The lack of suitable data makes the inference problem to be ill conditioned. A crucial issue is then how well the parameters need to be known in order to answer a given biological question. However, in all cases, it is important to validate the obtained estimates.

Sensitivity analysis can complement as well as support the parameter estimation (Saltelli et al., 2004; Fröhlich et al., 2016). In particular, the parameters can be ranked in the order of their importance, from the most easy to the most difficult to estimate. The parameters can be screened using a small amount of observations to select those which are identifiable prior to their inference from a full set of data. Other tasks in sensitivity analysis include prioritizing the parameters, testing their independence, and fixing or identifying the important regions of their values. A survey of methods used for the sensitivity analysis in BRNs is provided in Saltelli et al. (2005). The sensitivity profiles of 180 biological models were compared and analyzed in Erguler and Stumpf (2011).

In the rest of this paper, our main objective is to survey the models and methods which have been used in the literature to perform the parameter and state inferences in BRNs. After explaining our methodology in Section 2, different modeling strategies for BRNs are outlined in Section 3. It is followed by a survey of the estimation methods for BRNs and the related computational tasks in Section 4. Since the performance and effectiveness of estimation methods is crucially dependent on the specific models adopted, in Section 5, we explore what methods are used in literature for given models, and also, what estimation methods are used in given tasks. This enables us to uncover the possible future research directions in sub-section 5.1. We also mention several inference techniques which are used in other fields, but which can likely be assumed for BRNs.

Our contributions are 3-fold, and they are structured as the following surveys:

1. Models and modeling strategies of BRNs;
2. Parameter estimation methods, strategies and related tasks for models of BRNs;
3. Combinations of models and parameter estimation methods and tasks for BRNs.

The first version of this review appeared online as Loskot et al. (2019).

2. METHODOLOGY

It is important to define first the scope of our comprehensive review in order to understand its aims and constraints. In particular, there are at least 14 types of literature reviews

which differ in their purpose, methodology, and limitations (Grant and Booth, 2009). For example, the literature review can be systematic (SLR) to a various degree (Tranfield et al., 2003). The purpose of SLR is to answer an *a priori* formulated question or hypothesis using a clearly defined procedure of searching and examining the literature, so that it can be reproduced by others. The SLRs are particularly suitable for the evidence (data) based research fields as in biology and medicine (Grant and Booth, 2009).

However, the main purpose of our review is to present a comprehensive and critical overview of the models and methods which have been popular in literature to perform different inference tasks in BRNs. Such a review is known as the traditional or narrative literature review (NLR) (Onwuegbuzie and Frels, 2016). The outcome of NLR is state-of-the art of current knowledge, and identifying knowledge gaps, patterns, and emerging trends which can guide future research. The present review is comprehensive in the sense of striving to collect and categorize as many models and methods for inferences in BRNs as possible in order to provide a reference for further research on this topic. It leaves out the requirement for the review to be systematic and reproducible. We also cannot guarantee that all important and relevant papers in the field were identified or considered.

Our review resumed by collecting a relatively large number of representative and otherwise relevant papers. The papers were first identified using various keyword searches in Google. The subsequent more refined searches were performed in Google Scholar which also provides information on the citing papers, and contains the collections of papers by individual authors. Our intention was to specifically consider the papers on inference problems in BRNs; there are many other papers which are concerned with methods and strategies for general dynamic systems. We have also considered a number of graduate research theses which are publicly accessible online. The theses were evaluated separately from the papers. Moreover, we decided to exclude electronic books and textbooks from our study as their coverage is normally rather broad, and their contents processing would require to identify and extract chapters into separate files.

Almost 700 electronic documents in the portable document format (PDF) were collected from various sources using the following search keywords and their combinations: *biochemical*, *network*, *model*, *inference*, *estimation*, *parameters*, and *identification*. The initially collected papers were manually evaluated whether they are sufficiently relevant to the purpose of our study. For example, many papers involving parameter estimation in general dynamic systems were discarded unless they were deemed to have some other value for our review. While evaluating the papers, we were updating 2 lists of keywords. The first list contains keywords representing the models of BRNs, such as *state-space*, *differential equation*, *Markov chain*, and similar. The second keyword list describes the inference methods, for example, *Bayesian*, *MCMC*, *least squares*, and other. The keywords were used to perform more focused searches for additional papers, and to screen and classify the already collected papers. In the end, we assumed 25 BRN models and 23 inference methods, and also defined the 5 inference-related tasks: *estimation*, *inference*, *identifiability*, *observability*, *reachability*, *experiment design*, *bifurcation analysis*, and *sensitivity analysis*.

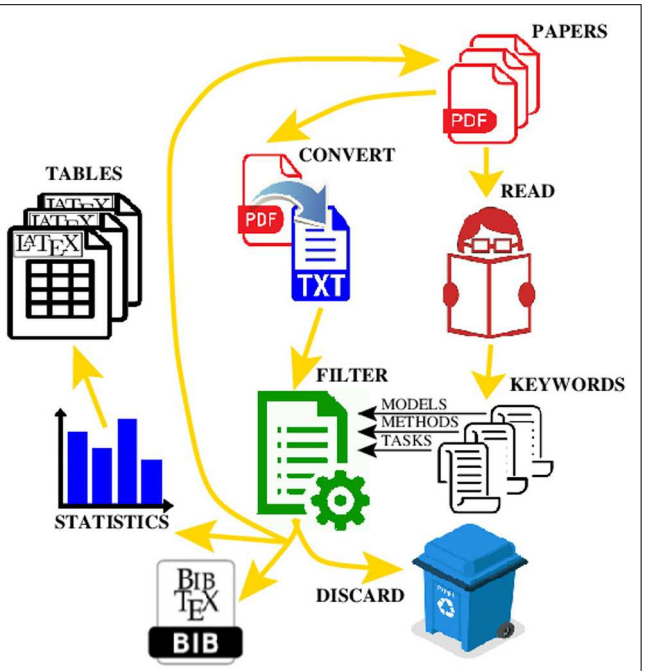


FIGURE 1 | A workflow for processing the PDF files to automate the production of the BIBTEX reference file, and the LATEX tables with statistical data.

All PDF documents were converted into ordinary text files to enable text mining of their contents. The text files were scanned to find occurrences of the keywords from the 2 lists defined above using the regular expressions representing textual patterns. The papers containing sufficiently large number of keywords were kept whereas the papers that did not pass the test were manually checked before being discarded. It allowed us to quickly reduce the number of papers from 700 to <300. There is a trade-off between the strictness (i.e., reliability) of the automated paper selection and possibility to automatically discard some papers, and how many of the remaining papers have to be checked manually. We observed that a small number of occurrences of a keyword usually indicates that the keyword appeared mainly within the references of the paper. A high-level view of our paper selection process is depicted in **Figure 1**.

As the number of published papers is increasing exponentially, there is clearly a need to develop new tools to facilitate more automated paper selection and pre-screening (Loskot, 2018). In order to automate many text processing tasks and enable evaluation of the 100's of papers in our study, we took advantage of the text processing capabilities readily available on the Linux operating systems. In particular, all PDF files were first converted to ordinary text files with the ascii encoding of characters (UTF-8) and the transliterated special characters in the foreign alphabets. The conversion was done using the

standard pdftotext utility version 0.62 which is based on the open source Poppler library developed for rendering the PDF files. The PDF conversion is not and does not have to be 100% accurate. For example, the words containing characters which are not recognized can be omitted. Moreover, some words are occasionally split into several parts which can be detected using a dictionary. However, such undesirable cases can be largely neglected for our purposes. It is also useful to remove the end-of-line characters from within the paragraphs, and to merge parts of the paragraphs which were split by displayed equations or by page breaks in order to improve the searches for more complex text patterns.

The scripts to automate many text processing tasks were programmed in the BASH interpreter version 4.4 running in a Linux terminal. The scripts use extensively standard Linux tools including grep, sed, and awk programmable text filters. In particular, the scripts were used to automatically identify and count relevant papers, generate LaTeX tables to visualize the results, facilitate semi-automated creation of bibliographic entries in the master BibTeX file, and to obtain URL links for citing papers in Google Scholar (**Table S3**). The keyword searches can assume multiple terms combined in sophisticated hierarchical expressions with AND-OR operators, include conditions on the number of occurrences, and sort the results as required.

However, the adopted procedure and the tools we developed for identifying and selecting the most relevant papers have some limitations. In particular, the paper selection and text mining in our study is restricted to keyword searches using regular expressions. A certain level of manual processing is still required, although it is likely that this can be reduced with future versions of the tools. A fully automated paper analysis with minimum human interventions would require the use of natural language processing (NLP). The NLP libraries are already available in many programming languages, but it is outside the scope of the present paper.

Furthermore, our study is mostly concerned with inferences of parameters and states whereas the inferences of network structures (i.e., which chemical reactions are occurring) is omitted. Our classification of models and methods have been developed to facilitate the analysis of trends and patterns in the literature. For instance, some models and methods considered in the next sections may be related, or a special case of one another. However, for the purpose of our study, the models and methods are presented as they appear in the cited references. In addition, although we generally distinguish between the deterministic and stochastic inference methods, we do not make such strict distinction between the deterministic and stochastic models. It should be also noted that many references can be cited in multiple contexts, i.e., for several models or methods considered. In many cases, the papers are chosen as illustrative examples for a given model or method, so they are likely many other important references which could be cited. Finally, more complete information how the papers cited in this review are related to the assumed models and methods is given in the **Supplementary Tables**.

3. REVIEW OF MODELING STRATEGIES FOR BRNS

Mathematical models describe dependencies of observations on the model parameters. A general procedure for constructing mathematical models of biological systems is described in Chou and Voit (2009). The bio-reactors are mathematically described in Vargas et al. (2014), Ali et al. (2015), and Farza et al. (2016). The model building is an iterative process which is often combined with the optimum experiment design (Rodriguez-Fernandez et al., 2006b). The model structure affects the selection as well as the performance of parameter estimators. The structural identifiability and validity of multiple models together with the parameter sensitivity was considered in Jaqaman and Danuser (2006). The parameter estimation can be performed together with the discrimination among several competing models, for instance, when the model structure is only partially known. The model structure and the parameter values to achieve the desired dynamics can be obtained by the means of statistical inference (Barnes et al., 2011). The synthesis of parameter values for BRNs is also considered in Češka et al. (2017). The probabilistic model checking can be used to facilitate the robustness analysis of stochastic biochemical models (Češka et al., 2014). The model checking is investigated in a number of references including Palmisano (2010), Brim et al. (2013), Češka et al. (2014), Mizera et al. (2014), Hussain et al. (2015), Mancini et al. (2015), Češka et al. (2017), and Milios et al. (2018). An iterative, feedback dependent modularization of models with the parameters identification was devised in Lang and Stelling (2016). A selection among several hierarchical models assuming Akaike information was studied in Rodriguez-Fernandez et al. (2013).

Modeling strategies of BRNs often involve the kinetics of chemical reactants which are described by the law of mass action or by the rate law (Schnoerr et al., 2017). Both these laws model the dependency of chemical reaction rates on the species concentrations. The reaction kinetics can be considered at steady state or in the transition to steady state, although the steady state may not be always achieved. There are also other kinetic models, such as the Michaelis-Menten kinetics for the enzyme-substrate reactions (Rumschinski et al., 2010), the Hill kinetics for cooperative ligand binding to macromolecules (Fey and Bullinger, 2010), the kinetics for logistic growth models in GRNs (Ghusinga et al., 2017), the kinetics for the birth-death processes (Daigle et al., 2012), and the stochastic Lotka-Volterra kinetics which are associated with the prey-predatory networks (Boys et al., 2008).

Single molecule stochastic models describe BRNs qualitatively by generating the probabilistic trajectories of species counts. A BRN can be modeled as a sequence of reactions occurring at random time instances (Amrein and Künsch, 2012). The stochastic kinetics mathematically correspond to a Markov jump process with the random state transitions between the species counts (Andreychenko et al., 2012). Alternatively, the time sequence of chemical reactions can be viewed as a hidden Markov process (Reinker et al., 2006). The Markov jump processes can be simulated exactly using the classical Gillespie algorithm, so

that the competing reactions are selected assuming a Poisson process with the intensity proportional to the species counts (Golightly et al., 2012; Kügler, 2012), although, in general, the intensity can be an arbitrary function of the species counts. The random occurrences of reactions can be also described using the hazard function (Boys et al., 2008). Non-homogeneous Poisson processes can be simulated by the thinning algorithm of Lewis and Shedler (Sherlock et al., 2014).

The number of species in BRN and their molecule counts can be large, so the state space of the corresponding continuous time Markov chain (CTMC) model is huge (Angius and Horváth, 2011). The large state space can be truncated by considering only the states significantly contributing to the parameter likelihood (Singh and Hahn, 2005). The parameter likelihoods can be updated iteratively assuming the increments and decrements of the species counts (Lecca et al., 2009). The probabilistic state space representations of BRNs as dynamic systems were considered in Andreychenko et al. (2011), Gupta and Rawlings (2014), McGoff et al. (2015), and Schnoerr et al. (2017). An augmented state space representation of BRN derived from the ordinary differential equations (ODEs) is obtained in Baker et al. (2013).

More generally, mechanistic models of BRNs are obtained by assuming that biological systems are built up from the actual or perceived components which are governed by the physical laws (Hasenauer, 2013; Pullen and Morris, 2014; White et al., 2016; Fröhlich et al., 2017). It is a different strategy to empirical models which are reverse-engineered from observations (Geffen et al., 2008; Bronstein et al., 2015; Dattner, 2015). The black-box modeling can be assumed with some limitations when there is little knowledge about the underlying biological processes (Chou and Voit, 2009).

Many models containing multiple unknown parameters are often poorly constrained. Even though such models may be still fully identifiable, they are usually ill-conditioned, and often referred to as being sloppy (Toni and Stumpf, 2010; Erguler and Stumpf, 2011; White et al., 2016). The parameter estimation and experimental design for sloppy models are investigated in Mannakee et al. (2016) where it is shown that the dynamic properties of sloppy models usually depend only on several key parameters with the remaining parameters being largely unimportant. A sequence of hierarchical models with increasing complexity was proposed in White et al. (2016) to overcome the complexity and sloppiness of conventional models.

3.1. Modeling BRNs by Differential Equations

The time evolution of states with the probabilistic transitions is described by a chemical master equation (CME) (Andreychenko et al., 2011; Weber and Frey, 2017). The CME is a set of coupled first-order ODEs or partial differential equations (PDEs) (Fearnhead et al., 2014; Penas et al., 2017; Teijeiro et al., 2017) representing a continuous time approximation and describing the BRN quantitatively. The ODE model of a BRN can be also derived as a low-order moment approximation of the CME (Bogomolov et al., 2015). For the models with stochastic differential equations (SDEs), it is often difficult to find the transition probabilities (Karimi and Mcauley, 2013; Fearnhead et al., 2014; Sherlock et al., 2014). The PDE approximation can be obtained assuming a Taylor expansion of the CME (Schnoerr et al., 2017). The error bounds for the numerically obtained stationary distributions of the CME are obtained in Kuntz et al. (2017). The CME for a hierarchical BRN consisting of the dependent and independent sub-networks is solved analytically in Reis et al. (2018). A path integral form of the ODEs has been considered in Liu and Gunawan (2014) and Weber and Frey (2017). The BRN models with memory described by the delay differential equations (DDEs) are investigated in Zhan et al. (2014). The mixed-effect models assume multiple instances of the SDE based models to evaluate statistical variations between and within these models (Whitaker et al., 2017).

A comprehensive tutorial on the ODE modeling of biological systems is provided in Gratie et al. (2013). The ODE models can be solved numerically via discretization. For instance, the finite differences method (FDM) can be used to obtain difference equations (Fröhlich et al., 2016). However, the algorithms for numerically solving the deterministic ODE models or simulating the models with SDEs may not be easily parallelizable, and they may have problems with numerical stability. The ODE models are said to be stiff, if they are difficult to solve or simulate, for example, if they comprise multiple processes at largely different time scales (Sun et al., 2012; Cazzaniga et al., 2015; Kulikov and Kulikova, 2017). Alternatively, the BRN structure can be derived from its ODE representation (Fages et al., 2015). A similar strategy is assumed in Plesa et al. (2017) where the BRN is inferred from the deterministic ODE representation of the time series data.

A survey of methods for solving the CME of gene expression circuits is provided in Veerman et al. (2018). These methods involve propagators, time-scale separation, and the generating functions (Schnoerr et al., 2017). For instance, the time-scale separation can be used to robustly decompose the CME into a hierarchy of models (Radulescu et al., 2012). A reduced stochastic description of BRNs exploiting the time-scale separation is studied in Thomas et al. (2012).

If the deterministic ODEs cannot be solved analytically, one can use Langevin and Fokker-Planck equations as the stochastic diffusion approximations of the CME (Hasenauer, 2013; Schnoerr et al., 2017). The Fokker-Planck equation can be solved to obtain a deterministic time evolution of the system state distribution (Kügler, 2012; Liao et al., 2015; Schnoerr et al., 2017). The deterministic and stochastic diffusion approximations of stochastic kinetics are reviewed in Mozgunov et al. (2018). The chemical Langevin equation (CLE) is a SDE consisting of a deterministic part describing the slow macroscopic changes, and a stochastic part representing the fast microscopic changes which are dependent on the size of the deterministic part (Golightly et al., 2012; Cseke et al., 2016; Dey et al., 2018). In the limit, as the deterministic part increases, the random fluctuations can be neglected, and the deterministic kinetics described by the Langevin equation becomes the reaction rate equation (RRE) (Bronstein et al., 2015; Fröhlich et al., 2016; Loos et al., 2016).

3.2. Modeling BRNs by Approximations

A popular strategy to obtain computationally efficient models is to assume approximations, such as meta-heuristics and meta-modeling (Sun et al., 2012; Cedersund et al., 2016). The quasi-steady state (QSS) and quasi-equilibrium (QE) approximations of BRNs are investigated in Radulescu et al. (2012). The modifications of QSS models are investigated in Wong et al. (2015). It is also common to approximate the system dynamics assuming continuous ODEs or SDEs (Fearnhead et al., 2014). The SDE model is preferred when the number of molecules is small, since the deterministic ODE model may be inaccurate (Gillespie and Golightly, 2012). It is generally difficult to quantify the approximation errors in the diffusion-based models. The forward-reverse stochastic diffusion with the deterministic approximation of propensities by the observed data was considered in Bayer et al. (2016).

The mass action kinetics can be used to obtain a deterministic approximation of CME. The corresponding deterministic ODEs can accurately describe the system dynamics, provided that the molecule counts of all the species are sufficiently large (Sherlock et al., 2014; Yenkie et al., 2016). Other CME approximations assume the finite state projections, the system size expansion, and the moment closure methods (Chevaliera and Samadb, 2011; Schnoerr et al., 2017). These methods are attractive, since they are easy to implement and efficient computationally. They do not require the complete statistical description, and they achieve good accuracy if the species appear in large copy numbers (Schnoerr et al., 2017). The moment closure methods leading to the coupled ODEs can approach the CME solution with a low computational complexity (Bogomolov et al., 2015; Fröhlich et al., 2016; Schilling et al., 2016). Specifically, the n-th moment of the population size depends on its $(n + 1)$ moment, and to close the model, the $(n + 1)$-th moment is approximated by a function of the lower moments (Ruess et al., 2011; Ghusinga et al., 2017). Only the first several moments can be used to approximate the deterministic solution of CME (Schnoerr et al., 2017). The limitations of the moment closure methods are analyzed in Bronstein and Koeppl (2018). A multivariate moment closure method is developed in Lakatos et al. (2015) to describe the non-linear dynamics of stochastic kinetics. The general moment expansion method for stochastic kinetics is derived in Ale et al. (2013). The approximations of the state probabilities by their statistical moments can be used to conduct efficient simulations of stochastic kinetics (Andreychenko et al., 2015).

The leading term of the CME approximation in the system size expansion (SSE) method corresponds to a linear noise approximation (LNA). It is the first order Taylor expansion of the deterministic CME with a stochastic component where the transition probabilities are additive Gaussian noises. Other terms of the Taylor expansion can be included in order to improve the modeling accuracy (Fröhlich et al., 2016). In Sherlock et al. (2014), the LNA is used to approximate the fast chemical reactions as a continuous time Markov process (CTMP) whereas the slow reactions are represented as a Markov jump process with the time-varying hazards. There are other variants of the LNA, such as a restarting LNA model (Fearnhead et al., 2014), the LNA with time integrated observations (Folia and Rattray, 2018), and the LNA with time-scale separation (Thomas et al., 2012). The LNA for the reaction-diffusion master equation (RDME) is computed in Lötstedt (2018). The impact of parameter values on the stochastic fluctuations in a LNA of BRN is investigated in Pahle et al. (2012).

The so-called S-system model is a set of decoupled non-linear ODEs in the form of product of power-law functions (Chou et al., 2006; Meskin et al., 2011; Liu et al., 2012; Iwata et al., 2014). Such models are justified by assuming a multivariate linearization in the logarithmic coordinates. These models provide a good trade-off between the flexibility and accuracy, and offer other properties which are particularly suitable for modeling complex non-linear systems. The S-system models with additional constraints are assumed in Sun et al. (2012). The S-system modeling of biological pathways is investigated in Mansouri et al. (2015). The S-system model with weighted kinetic orders is obtained in Liu and Wang (2008a). The Bayesian inference for S-system models is investigated in Mansouri et al. (2014).

Polynomial models of biological systems are investigated in Kuepfer et al. (2007), Vrettas et al. (2011), Fey and Bullinger (2010), and Dattner (2015). Rational models as fractions of polynomial functions are examined in Fey and Bullinger (2010), Eisenberg and Hayashi (2014), and Villaverde et al. (2016). The methods for validating polynomial and rational models of BRNs are studied in Rumschinski et al. (2010). The eigenvalues are used in Hori et al. (2013) to obtain a low order linear approximation of the time series data. More generally, the models with differential-algebraic equations (DAEs) are considered in Ashyraliyev et al. (2009), Michalik et al. (2009), Rodriguez-Fernandez et al. (2013), and Deng and Tian (2014). These models have different characteristics than the ODE based models, and they are also more difficult to solve. The review of autoregressive models for parameter inferences including the stability and causality issues is presented in Michailidis and dAlchéBuc (2013).

3.3. Other Models of BRNs

There are many other types of BRN models considered in the literature. The birth-death process is a special case of the CTMP having only two states (Daigle et al., 2012; Paul, 2014; Zechner, 2014). It is closely related to a telegraph process (Veerman et al., 2018). A computationally efficient tensor representation of BRNs to facilitate the parameter estimation and sensitivity analysis is devised in Liao et al. (2015). Other computational models for a qualitative description of interactions and behavioral logic in BRNs involve the Petri nets (Mazur, 2012; Sun et al., 2012; Schnoerr et al., 2017), the probabilistic Boolean networks (Liu et al., 2012; Mazur, 2012; Mizera et al., 2014), the continuous time recurrent neural networks (Berrones et al., 2016), and the agent based models (ABMs) (Hussain et al., 2015). The hardware description language (HDL) originally devised to describe the logic of electronic circuits is adopted in Rosati et al. (2018) to model spatially-dependent biological systems with the PDEs. The multi-parameter space was mapped onto a 1D manifold in Zimmer et al. (2014).

TABLE 1 | An overview of the main modeling strategies for BRNs.

Strategy	Motivation and key papers
Physical laws	Reaction rates in dynamic equilibrium are functions of reactant concentrations
• Kinetic rate laws	Joshia et al., 2006; Chou and Voit, 2009; Engl et al., 2009; Baker et al., 2011; Villaverde et al., 2012; Voit, 2013
• Mass action kinetics	Angius and Horváth, 2011; Lindera and Rempala, 2015; Wong et al., 2015; Smith and Grima, 2018
• Mechanistic models	Chou and Voit, 2009; Pullen and Morris, 2014; von Stosch et al., 2014; White et al., 2016
Random processes	Probabilistic behavioral description of chemical reactions
• Markov process	Andrieu et al., 2010; Goutsias and Jenkinson, 2013; Septier and Peters, 2016; Weber and Frey, 2017
• Poisson process	Daigle et al., 2012; Weber and Frey, 2017; Bronstein and Koeppl, 2018; Reis et al., 2018
• Birth-death process	Wang et al., 2010; Daigle et al., 2012; Mikelson and Khammash, 2016; Weber and Frey, 2017
• Telegraph process	Weber and Frey, 2017; Veerman et al., 2018
Mathematical models	Adopted models for dynamic systems
• Quasi-state models	Radulescu et al., 2012; Srivastava, 2012; Thomas et al., 2012; Wong et al., 2015; Liao, 2017; Schnoerr et al., 2017
• State space representation	Andrieu et al., 2010; Andreychenko et al., 2011; Brim et al., 2013; Weber and Frey, 2017
• ODEs, PDEs, SDEs, DDEs	J. O. Ramsay and Cao, 2007; Jia et al., 2011; Liu and Gunawan, 2014; Fages et al., 2015; Teijeiro et al., 2017; Weber and Frey, 2017
• Path integral form of ODEs	Weber and Frey, 2017
• Rational model	Sun et al., 2012; Vanlier et al., 2013; Hussain et al., 2015; Villaverde et al., 2016
• Differential algebraic eqns.	J. O. Ramsay and Cao, 2007; Ashyraliyev et al., 2009; Michalik et al., 2009; Deng and Tian, 2014
• Tensor representation	Liao et al., 2015; Wong et al., 2015; Smith and Grima, 2018
• S-system model	Kutalik et al., 2007; Chou and Voit, 2009; Meskin et al., 2011; Liu et al., 2012; Voit, 2013
• Polynomial model	Vrettas et al., 2011; Češka et al., 2017; Kuntz et al., 2017; Weber and Frey, 2017
• Manifold map	Radulescu et al., 2012; Mannakee et al., 2016; Septier and Peters, 2016; White et al., 2016
Interaction models	Qualitative modeling of chemical interactions
• Petri nets	Chou and Voit, 2009; Liu et al., 2012; Voit, 2013
• Boolean networks	Chou and Voit, 2009; Emmert-Streib et al., 2012
• Neural networks	Goutsias and Jenkinson, 2013; von Stosch et al., 2014; Ali et al., 2015; Camacho et al., 2018
• Agent based models	Carmi et al., 2013; Goutsias and Jenkinson, 2013; Hussain et al., 2015; Jagiella et al., 2017
CME based models	Stochastic and deterministic approximations of CME
• Langevin equation	Thomas et al., 2012; Goutsias and Jenkinson, 2013; Septier and Peters, 2016; Schnoerr et al., 2017; Weber and Frey, 2017; Smith and Grima, 2018
• Fokker-Planck equation	Liao et al., 2015; Schnoerr et al., 2017; Weber and Frey, 2017
• Reaction rate equation	Koeppl et al., 2012; Liu and Gunawan, 2014; Lindera and Rempala, 2015; Loos et al., 2016
• Moment closure	Ruess et al., 2011; Andreychenko et al., 2015; Lakatos et al., 2015; Schilling et al., 2016; Schnoerr et al., 2017; Bronstein and Koeppl, 2018
• Linear noise approximation	Golightly et al., 2012, 2015; Thomas et al., 2012; Fearnhead et al., 2014; Schnoerr et al., 2017; Whitaker et al., 2017
• System size expansion	Fröhlich et al., 2016; Schnoerr et al., 2017

The hybrid models generally combine different modeling strategies in order to mitigate various drawbacks of specific strategies (Mikeev and Wolf, 2012; Sherlock et al., 2014; Babtie and Stumpf, 2017). For example, a hybrid model can assume deterministic description of large species populations with the stochastic variations of small populations (Mikeev and Wolf, 2012). The hybrid model consisting of the parametric and non-parametric sub-models can offer some advantages over mechanistic models (von Stosch et al., 2014).

The modeling strategies discussed in this section are summarized in **Table 1**. The models are loosely categorized as physical laws, random processes, mathematical models, interaction models and the CME based models. These models are mostly quantitative except the interaction based models which are qualitative. Note that the model properties, such as sloppiness, and the model structures which may be hierarchical, modular or sequential are not distinguished in **Table 1**.

In order to assess the level of interest in different BRN models in literature, **Table S1** presents the number of occurrences for the 25 selected modeling strategies in all references cited in this review. The summary of **Table S1** is reproduced in **Table 2** with the inserted bar graph, and further visualized as a word cloud in **Figure 2**. We observe that differential equations are the most commonly assumed models of BRNs in the literature. About half of the papers cited consider the Markov chain models or their variants, since these models naturally and accurately represent

TABLE 2 | The coverage of modeling strategies for BRNs.

	Physical laws			Random processes				Mathematical models								Interaction models				CME based models					
	Kinetic rate laws	Mass action kinetics	Mechanistic models	Markov process	Poisson process	Birth-death process	Telegraph process	State space representation	ODEs, PDEs, SDEs, DDEs	Rational model	Differential algebraic eqns	Tensor representation	S-system model	polynomial model	manifold map	Petri nets	Boolean networks	Neural networks	Agent based models	Langevin equation	Fokker-Planck equation	Reaction rate equation	Moment closure	Linear noise approximation	System size expansion
# papers	63	102	84	170	74	21	2	152	219	60	27	21	39	90	35	13	13	37	47	55	35	19	46	51	4

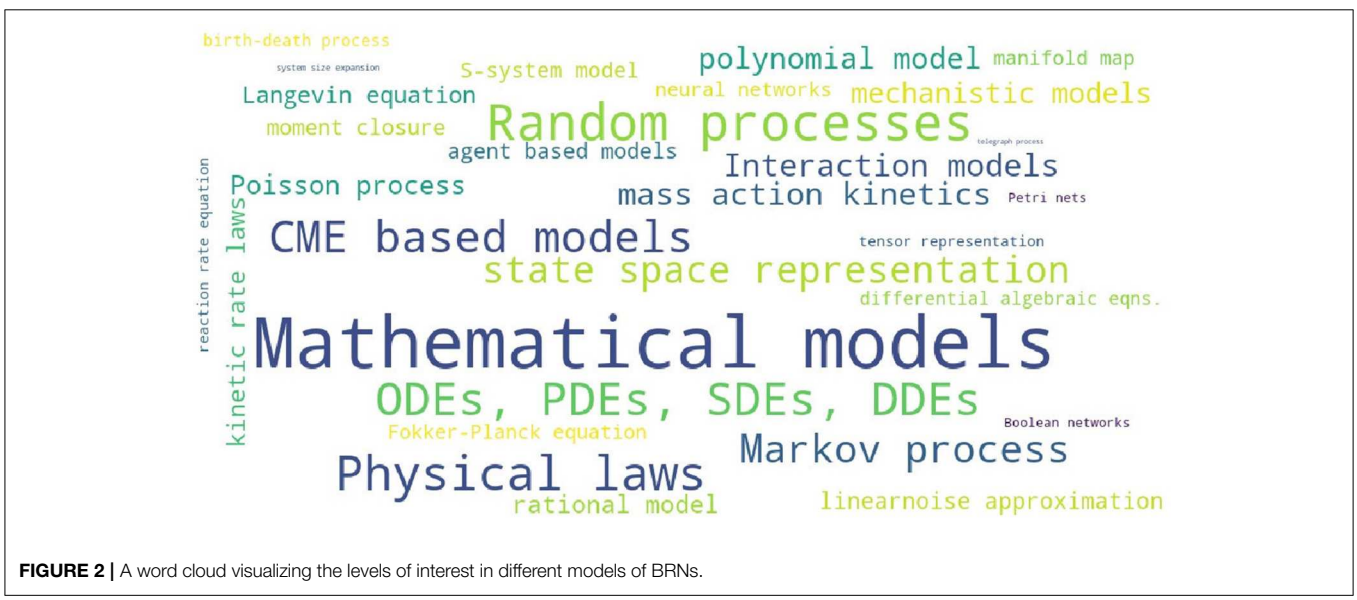


FIGURE 2 | A word cloud visualizing the levels of interest in different models of BRNs.

the time sequences of randomly occurring reactions in BRNs. The state space representations are assumed in over one third of the cited papers. Other more common models of BRNs include the mass action kinetics, mechanistic models, and the models involving polynomial functions.

Another viewpoint on BRN models in literature is to consider the publication years of papers. **Table 3** shows the number of papers for a given modeling strategy in a given year starting from the year 2005. The dot values in tables represent zero counts to improve the readability. We can observe that the interest in some modeling strategies remain stable over the whole decade, for example, for the models involving state space representations and the models involving differential equations. The number of cited papers is the largest in years 2013 and 2014. The paper counts in **Table 3** indicate that the interest in computational modeling of BRNs has been increasing steadily over the past decade.

4. REVIEW OF PARAMETER ESTIMATION STRATEGIES FOR BRNS

The parameter estimation or inference appears in many other computational problems including model identification (Banga and Canto, 2008), model calibration (Zechner et al., 2011), model discrimination (Kuepfer et al., 2007), model identifiability (Geffen et al., 2008), model checking (Hussain et al., 2015), sensitivity analysis (Erguler and Stumpf, 2011), optimum experiment design (Ruess and Lygeros, 2015), bifurcation

TABLE 3 | The number of papers concerning models of BRNs in given years.

	Physical laws			Random processes				Mathematical models								Interaction models				CME based models						
Year	Kinetic rate laws	Mass action kinetics	Mechanistic models	Markov process	Poisson process	Birth-death process	Telegraph process	State space representation	ODEs, PDEs, SDEs, DDEs	Rational model	Differential algebraic eqns.	Tensor representation	S-system model	Polynomial model	Manifold map	Petri nets	Boolean networks	Neural networks	Agent based models	Langevin equation	Fokker-Planck equation	Reaction rate equation	Moment closure	Linear noise approximation	System size expansion
2005	3	3	1	2	2	.	.	4	5	.	2	.	.	.	.	.	.	.	.	1	2	.	.	.	.
2006	2	4	2	2	.	.	.	2	3	4	1	.	1	3	.	.	.	2	.	.	.	.	.	.	.
2007	.	4	1	3	2	.	.	2	6	1	1	.	2	4	2	.	.	1	1	1	.	.	.	.	.
2008	1	4	2	2	1	.	.	6	6	1	1	.	2	2	1	.	.	2	.	.	.	1	.	.	.
2009	4	7	2	5	1	.	.	6	11	1	2	1	4	2	2	2	2	2	3	1	1	.	.	1	.
2010	7	11	5	12	3	1	.	8	13	6	2	.	4	5	5	1	1	2	5	7	2	1	.	2	.
2011	5	4	5	11	4	.	.	10	13	2	.	.	3	4	1	.	.	2	2	4	2	2	2	2	.
2012	6	11	6	22	11	3	.	15	20	6	4	1	5	6	3	2	3	6	7	9	6	1	4	9	.
2013	7	9	12	16	8	3	.	17	26	9	3	1	7	12	4	2	.	3	7	6	2	2	4	4	.
2014	8	13	14	33	11	4	.	26	33	7	4	2	5	14	2	1	3	7	6	7	5	5	10	9	.
2015	7	10	8	15	5	2	.	20	24	5	1	2	3	10	4	2	1	1	3	4	4	.	6	5	.
2016	4	8	13	20	5	2	.	14	23	4	1	2	3	11	5	1	2	5	7	6	3	3	7	8	2
2017	4	8	8	14	11	6	1	12	18	7	4	8	.	10	7	1	1	4	4	6	5	2	8	5	2
2018	4	8	4	11	8	1	1	7	14	5	.	3	.	5	.	1	.	1	2	4	3	2	4	5	.

analysis (Engl et al., 2009), reachability analysis (Tenazinha and Vinga, 2011), causality analysis (Carmi et al., 2013), stability analysis (Dochain, 2003), network inference (Smet and Marchal, 2010), and network control (Venayak et al., 2018). A chemical reaction optimization (CRO) can be used to maximize the production of a bio-reactor (Abdullah et al., 2013b). The surveys of parameter estimation methods for chemical reaction systems can be found, for example, in Chou and Voit (2009), Gupta (2013), Baker et al. (2015), and McGoff et al. (2015). Other review papers on parameter estimation in BRNs and dynamic systems are listed in **Table 4**.

A survey of tasks concerning modeling and system identification is provided in Chou and Voit (2009). The model identifiability determines which parameters can be estimated from observations (Villaverde et al., 2016). It is inspired by the concept of system observability and known as a structural identifiability. It is useful to consider the structural identifiability prior to estimating the parameters. There is also a practical identifiability which accounts for the quality and quantity of observations, i.e., whether it is possible to obtain good parameter estimates from noisy and limited data. The theory and tools for the model identifiability and other closely related concepts, such as the sensitivity to parameter perturbations, the observability, the distinguishability and the optimum experiment design are reviewed in Villaverde and Barreiro (2016). The models which are not identifiable can be modified or simplified to make them identifiable (Baker et al., 2015; Villaverde and Barreiro, 2016; Villaverde et al., 2016). The model identifiability is formulated as the model observability in Geffen et al. (2008) by replacing traditional analytical approaches which often require model simplifications with other deterministic empirical methods.

The changes in the structural and practical identifiability of models when new knowledge and data become available is studied in Babtie and Stumpf (2017). The global observability and detectability of reaction systems was studied in Moreno and Denis (2005). The parameter identifiability of the power law models is investigated in Srinath and Gunawan (2010) and of the linear dynamic models in Li and Vu (2013). The parameter dependencies are considered in Li and Vu (2015) to determine the structural and practical identifiability. The intrinsic noise in the species counts can be exploited to overcome the structural

TABLE 4 | The review papers on the parameter estimation in BRNs and other dynamic systems.

Reference	Focus
Banga and Canto, 2008	Model calibration using global optimization methods supported by maximum information experiment design
Chou and Voit, 2009	Very comprehensive survey of available optimization methods for parameter estimation and model-free and model-based structure identification from data
Ashyraliyev et al., 2009	*A priori* and a posteriori model identifiability and survey of parameter space search strategies
Smet and Marchal, 2010	Methods for under-determined inferences of BRNs from data
Tenazinha and Vinga, 2011	Integrated models of BRNs reflecting availability of omics data assuming chemical organization theory, flux-balance analysis, logical discrete modeling, Petri nets, kinetic models, stochastic models, and hybrid models
Daigle et al., 2012	Survey of maximum-likelihood based methods
Emmert-Streib et al., 2012	Systematic and conceptual overview of methods for inferring gene regulatory networks from gene expression data; survey of strategies to compare performance of inference methods
Sun et al., 2012	Survey of metaheuristic methods applied to reliability and identifiability of biochemical model parameters including optimum experiment design
Goutsias and Jenkinson, 2013	Comprehensive review of analytical methods for evaluating dynamics of Markov reaction networks
Kuwahara et al., 2013	Scalable framework for parameter estimation in genetic circuits assuming mean time evolution of gene products
Voit, 2013	Review of biological system models and methods for their analysis as well as design
Baker et al., 2015	General framework to deal with non-identifiable parameters in BRNs using constrained parameter estimation
McGoff et al., 2015	Mathematical survey of statistical methods for parameter inference in general non-linear dynamical systems
Drovandi et al., 2016	Survey of approximate Bayesian computation methods
Kurt et al., 2016	Review of 27 estimators of association scores of data from gene networks
Weiss et al., 2016	Survey of transfer learning methods
Schnoerr et al., 2017	A comprehensive survey of deterministic and stochastic models of BRNs followed by introduction to Bayesian parameter inference from data
Camacho et al., 2018	Application of machine learning techniques to computational problems in biological networks
Smith and Grima, 2018	Review of spatial stochastic kinetics including reaction-diffusion master equation and models involving Brownian dynamics
Koblents et al., 2019	Bayesian inference methods with stochastic kinetic models

non-identifiability within a deterministic framework as shown in Zimmer et al. (2014).

In general, many different parameter estimation methods have been devised in literature for BRNs and dynamic systems. However, many of these methods are often modifications of a few fundamental estimation strategies which are adopted for the specific models and the availability and quality of measurements. All parameter estimation problems lead to the minimization or maximization of some fitness function. Deriving the optimum value analytically is rarely possible whereas a numerical search for the optimum in high-dimensional parameter spaces can be ill-conditioned when the fitness function is multi-modal. The numerical strategies normally experience a trade-off between the efficiency and robustness. If there is a large flat surface about the minimum, the obtained solution cannot be trusted (Rodriguez-Fernandez et al., 2006a; Srinivas and Rangaiah, 2007). Moreover, the optimum values can change over an order of magnitude under different implicit or explicit constraints which is often the case for biological systems. The numerical algorithms for non-convex optimization problems need to be stable as well as provide the convergence guarantees. Other important aspects to consider include scalability, computational efficiency, numerical stability and robustness. All methods need to be also statistically validated.

The measurements can be produced from different heterogeneous sources (omics data), and from heterogeneous populations (Zechner et al., 2011). In literature, the deterministic methods appear to be assumed much more often than the stochastic methods (Daigle et al., 2012). The parameter estimation in deterministic models is often carried out by fitting the model to the data. The parameter uncertainty analysis can be used to assess how well the model explains the experimental data (Vanlier et al., 2013). The stochastic models require more sophisticated strategies to perform parameter estimation (Zimmer and Sahle, 2012), such as the multiple-shooting methods (Zimmer, 2016). Moreover, since the mean approximation of SDEs may differ from the solution obtained for deterministic ODEs, the parameter estimation assuming stochastic rather than deterministic models is preferable when some of the species counts are relatively small (Andreychenko et al., 2012).

The parameter estimations in the transient and at steady state are quite different (Ko et al., 2009). At steady state, small perturbations are sufficient to observe the system responses whereas at the transient state, the experiment design for model identification is more complicated. A fast transient response after the external perturbation limits the information content in measurements (Zechner et al., 2012). The sensitivity analysis can be used to improve the computational efficiency of parameter estimation (Fröhlich et al., 2017). The parameter value boundaries can be estimated by sampling (Fey and Bullinger, 2010). The confidence and credible intervals can be obtained also for the stiff and sloppy models assuming

the inferability, sensitivity and sloppiness (Erguler and Stumpf, 2011). Furthermore, the observer design may be different for systems with and without inputs (Singh and Hahn, 2005).

The scalability of parameter estimation can be resolved by decoupling the rate equations and by assuming the mean-time evolution of the species counts (Kuwahara et al., 2013). However, exploring large parameter spaces can be complicated, if the estimation problems are ill-conditioned and multi-modal (Liu and Wang, 2009). The state-dependent Markov jump processes are difficult to estimate at large scale, especially when these processes are faster than the rate of observations (Fearnhead et al., 2014).

The model parameters can be mutually dependent (Fey et al., 2008). The parameter dependencies can be measured by correlations and other higher order moments. The parameter estimation can be facilitated by grouping the parameters, and then identifying which are uncorrelated (Gábor et al., 2017). The parameter estimation in groups can provide robustness against the noisy and incomplete data (Jia et al., 2011). Only the parameters which are consistent with the measured data can be selected and jointly estimated (Hasenauer et al., 2010). The parameter clustering can also improve the model tractability and identifiability, since the changes in some parameters could be compensated by changes in other parameters (Nienaltowski et al., 2015). The groupings of parameters to elucidate the dynamics of genetic circuits are assumed in Atitey et al. (2019). The parameters can be assumed hierarchically to gradually estimate their values starting from a suitably defined minimum set (Shacham and Brauner, 2014). A hybrid hierarchical parameter estimation method which is prone to parallel implementation is devised in He et al. (2004).

An incremental parameter estimation usually requires data smoothing which can create the estimation biases (Liu and Gunawan, 2014). Such biases can be mitigated by estimating the independent parameters before the dependent ones. The parameter inference can be paired with the hypothesis testing and model selection (Rodriguez-Fernandez et al., 2013). The joint model and parameter identification with incremental one-at-a-time parameter estimation and model building is performed in Gennemark and Wedelin (2007). The unobserved states, latent variables and other parameters in BRNs can be estimated jointly by sequentially processing the measurements (Zimmer and Sahle, 2012; Arnold et al., 2014), by using the sliding window observers (Liu et al., 2006), and by other numerical methods (Karnaukhov et al., 2007). The estimation of kinetic rates in BRNs is transformed into a problem of the state estimation in Fey and Bullinger (2010). The parameter estimation and the state reconstruction are linked via the extended models in Busetto and Buhmann (2009). The unobservable sub-spaces can be excluded, and only the model parts which are identified reliably can be considered (Singh and Hahn, 2005). Another strategy is to reconstruct the states prior to estimating the parameters (Fey et al., 2008). The unknown parameters which are not of interest can be margninalized (Bronstein et al., 2015).

The model overfitting leads to a poor generalization capability. In order to avoid the overfitting and to constrain the model complexity, a penalty can be assumed to minimize the number of estimated model parameters. The overfitting can be resolved by the model reduction techniques (Srivastava, 2012; Sadamoto et al., 2017). For instance, only essential chemical reactions can be considered in BRN model (Zamora-Sillero et al., 2011). A simplified modeling with the reduced number of parameters and the parameter subset selection is used in Eghtesadi and Mcauley (2014) to avoid overfitting the noisy data. On the other hand, the under-determined models may yield several or infinitely many solutions of fitting the data. In such cases, the models are not identifiable, and the data fitting can be performed subject to additional constraints. There are also cases where the measured data can be fit well by several models. However, the model with the best fit to the data may not necessarily provide a satisfactory biological explanation (Slezak et al., 2010).

The information theoretic metrics can be used to infer the structure of BRNs (Villaverde et al., 2014), and to perform the identifiability analysis of parameters (Nienaltowski et al., 2015). Akaike information is used to assess the quality of statistical models given observations, so the best model can be selected (Guillén-Gosálbez et al., 2013; Pullen and Morris, 2014). The simultaneous estimation of parameters and the structure of BRN formulated as a mixed binary dynamic optimization problem with Akaike information is assumed in Guillén-Gosálbez et al. (2013) to trade-off the estimation accuracy and the evaluation complexity. Fisher information is the mean amount of information gained from the observed data. It is often used when estimating the non-random parameters, for instance, using the maximum likelihood (ML) (Rodriguez-Fernandez et al., 2006b; Kyriakopoulos and Wolf, 2015). Fisher information can be exploited to perform the sensitivity, robustness and identifiability of parameters. It is especially useful when the measurements and parameters are correlated (Komorowski et al., 2011). Fisher information is also used to improve the parameter estimation (Transtrum and Qiu, 2012), to design the optimum experiments (Kyriakopoulos and Wolf, 2015; Zimmer, 2016), and to select the subsets of identifiable parameters (Eisenberg and Hayashi, 2014). Mutual information can be used as a similarity measure. It statistically outperforms correlations in the canonical correlation analysis (CCA) (Nienaltowski et al., 2015). Other uses of mutual information are outlined in Mazur (2012), and for the parameter estimation in Emmert-Streib et al. (2012).

The cross-entropy methods can be combined with stochastic simulations (Revell and Zuliani, 2018), and used to improve the computational efficiency of the parameter estimation (Daigle et al., 2012). The maximum entropy sampling (MES) methods for the experiment design and for the parameter estimation are discussed in Mazur and Kaderali (2013). The maximum entropy principle to reconstruct the probability distributions is described in Schnoerr et al. (2017). The relative entropy rate is assumed in Pantazis et al. (2013) to perform the sensitivity analysis of BRNs. The Kantorovich distance between two probability measures is used in Koeppl et al. (2010) to estimate the BRN model parameters.

The sum of squared errors (SSE) is often assumed to define the regression estimators (Chou et al., 2006), to evaluate the goodness of fit, and to assess the quality of estimators (Nim et al., 2013; Iwata et al., 2014; Kimura et al., 2015). The SSE acronym should

TABLE 5 | The selected research theses concerning the parameter estimation and related problems in BRNs.

Thesis	Main research focus
Dargatz, 2010	Bayesian inference for biochemical models involving diffusion
Mu, 2010	Rate and state estimation in S-system and linear fractional model (LFM)
Palmisano, 2010	Software tools for modeling and parameter estimation in BRNs
Mazur, 2012	Inference via stochastic sampling and Bayesian learning framework
Srivastava, 2012	Stochastic simulations of BRNs combined with likelihood based parameter estimation, confidence intervals, sensitivity analysis
Gupta, 2013	Parameter estimation in deterministic and stochastic BRNs, inference with model reduction, mostly MCMC methods
Hasenauer, 2013	Bayesian estimation and uncertainty analysis of population heterogeneity and proliferation dynamics
Linder, 2013	Penalized LS algorithm and diffusion and linear noise approximations and algebraic statistical models
Flassig, 2014	Model identification for large scale gene regulatory networks
Liu, 2014	Approximate Bayesian inference methods and sensitivity analysis
Moritz, 2014	Structural identification and parameter estimation for modular and layered type of modes
Paul, 2014	Analysis of MCMC based methods
Ruess, 2014	Optimum estimation and experiment design assuming ML and Bayesian inference and Fisher information
Schenkendorf, 2014	Quantification of parameter uncertainty, optimal experiment design for parameter estimation and model selection
Smadbeck, 2014	Moment closure methods, model reduction, stability and spectral analysis of BRNs
Zechner, 2014	Inference from heterogeneous snapshot and time-lapse data
Schnoerr, 2016	Langevin equation, moment closure approximations, representations of stochastic RDME
Galagali, 2016	Bayesian and non-Bayesian inference in BRNs, adaptive MCMC methods, network-aware inference, inference for approximated BRNs
Hussain, 2016	Sequential probability ratio test, Bayesian model checking, automated and formal verification, parameter discovery
Lakatos, 2017	Multivariate moment closure and reachability analysis
Liao, 2017	Tensor representation and analysis of BRNs

not be confused with the system size expansion (SSE) which is a modeling strategy discussed previously (Fröhlich et al., 2016; Schnoerr et al., 2017).

Furthermore, the graduate research theses usually contain more or less comprehensive and up to date surveys of the relevant literature. The theses which are concerned with the parameter estimation in BRNs are summarized in **Table 5**. We can observe that the largest number of the research theses involving the parameter estimation problems in BRNs were produced in 2014.

In the rest of this section, we will survey specific methods for the parameter estimation in BRNs. These methods are organized in the following four subsections: Bayesian methods, Monte Carlo methods, other statistical methods including Kalman filtering, and the model fitting methods.

4.1. Bayesian Methods

The fundamental premise of the Bayesian estimation methods is that the prior probabilities or distributions of parameters are known. The objective is then to evaluate the posterior distributions for the parameters of interest. It is often sufficient to find the maximum value of the posterior distribution as the maximum a posterior (MAP) estimate. The value of this maximum can be also used to select among several competing models (Andreychenko et al., 2012) and to design the optimum experiments (Mazur, 2012). The model checking via the time-bounded path properties is represented as the Bayesian inference problem in Milios et al. (2018). The conjugate priors are often assumed in biological models to perform the Bayesian inferences (Boys et al., 2008; Mazur, 2012; Murakami, 2014; Galagali, 2016). The Bayesian inference for the low copy counts can be improved by separating the intrinsic and extrinsic noises (Koeppl et al., 2012). The Bayesian analysis is facilitated by separating the slow and fast reactions in Sherlock et al. (2014). The Bayesian inference strategies for biological models involving diffusion processes are investigated in Dargatz (2010).

In many cases, determining the exact posterior distribution in the Bayesian analysis is analytically intractable. The approximate Bayesian computation (ABC) is a computational strategy for estimating the posterior distribution or the likelihood function (Tanevski et al., 2010). The survey of ABC approaches is provided in Drovandi et al. (2016). The basic idea is to find the parameter values which can generate the same statistics as the observed data. The ABC can be performed sequentially, and used for the sensitivity analysis (Liu, 2014). The parameter estimation and the model selection using the ABC framework is studied in Liepe et al. (2014) and Murakami (2014). The non-identifiability of parameters due to the flat-shaped posterior can be resolved by the ABC approach as shown in Murakami (2014). The efficient generation of summary statistics for the ABC is presented in Fearnhead and Prangle (2012). The piece-wise ABC to estimate the posterior density for Markov models is proposed in White et al. (2015). The parallel implementations of the ABC and SMC methods are introduced in Jagiella et al. (2017).

The expectation-maximization (EM) is a popular implementation of the MAP estimators where there are some other unobserved or unknown parameters (Daigle et al., 2012; Karimi and Mcauley, 2014a; Bayer et al., 2016). The EM can be combined with the Monte Carlo (MC) sampling, and

such methods are known as the MC expectation-maximization (MCEM) (Angius and Horváth, 2011). The computationally efficient method for obtaining the ML estimates by the MCEM with a modified cross-entropy method (MCEM2) is developed in Daigle et al. (2012). The approximate EM algorithm is devised in Karimi and Mcauley (2013) which is robust against the unknown initial estimates, and which is useful for the online state estimation during the process monitoring.

Another parameter estimation strategy having the same structure as the EM is known as the variational Bayesian inference (Vrettas et al., 2011; Weber and Frey, 2017). It is more general than the EM method, and it exploits the analytical approximations of the posterior density to obtain the parameter estimates and their likelihoods. The analytical approximations are usually computationally faster than the sampling based methods, but the approximation methods are still less well-understood (Blei et al., 2017). For instance, the posterior density is approximated by radial basis functions (RBFs) in Fröhlich et al. (2014) to reduce the number of model evaluations. The variational inference with stochastic approximations for Gaussian mixture models and massive data is considered in Blei et al. (2017). The variational approximate inference with the continuous time constraints is investigated in Cseke et al. (2016).

The ML estimation is a popular parameter estimation strategy, provided that the likelihoods of the observed data can be computed efficiently for the given model. The survey of ML based methods for the parameter estimation in BRNs is provided in Daigle et al. (2012). The likelihood function can be approximated analytically using the Laplace and the B-spline approximations (Karimi and Mcauley, 2014b), or numerically by assuming the derivatives (Mikeev and Wolf, 2012). The likelihood function is obtained by simulations in Tian et al. (2007). The moment closure is used for the fast approximations of the parameter likelihoods in Milner et al. (2013). Stochastic simulations can be avoided by approximating the transition distributions by the Gaussian distribution in the parameter likelihood calculations (Zimmer and Sahle, 2015). In Chen et al. (2017), the transition probabilities are used in the ML calculations to devise the new estimation algorithm which can improve the variational Bayesian inference. The ML estimation combined with regularization to penalize the complexity is investigated in Jang et al. (2016). The ML estimation for BRN models with the concentration increments and decrements is studied in Lecca et al. (2009).

4.2. Monte Carlo Methods

The motivation behind the MC methods is to represent the probabilities and density functions as the relative frequencies of samples or particles in order to overcome mathematical intractability of the Bayesian inference. However, even the sampling methods can be computationally overwhelming due to frequent model evaluations. The Markov chain Monte Carlo (MCMC) methods are the most often used sampling strategies to generate conditional trajectories of the system states. The MCMC sampling having good mixing properties requires a carefully chosen proposal distribution and also a good selection of the initial samples in order to avoid the sample degeneracy and instability problems. The most well-known sampling MCMC procedures are the Metropolis and the Metropolis-Hastings algorithms (Golightly and Wilkinson, 2011; Zamora-Sillero et al., 2011; Mazur, 2012; Galagali, 2016). An overview of the particle filtering and the MCMC methods for the spatial objects tracking is presented in Mihaylova et al. (2014). The MCMC methods for causality reasoning are introduced in Carmi et al. (2013). The design of proposal distributions for the MCMC and the SMC methods assuming a large number of correlated variables is studied in Andrieu et al. (2010).

Since the convergence rate of the MCMC sampling can be rather slow for heavy tail distributions, the factorization and approximations of the posterior can be used to improve the performance (Fröhlich et al., 2014). The MCMC methods can be made adaptive to improve their convergence properties as shown in Mazur (2012); Müller et al. (2012); Hasenauer (2013); Galagali (2016). The interpolation of the observed data via the MCMC sampling is assumed in Golightly and Wilkinson (2005) to jointly estimate the unobserved states and reaction rates. The MCMC sampling can be combined with the importance sampling to reduce the computational complexity and simulation times (Golightly et al., 2015). The conditional density importance sampling (CDIS) is introduced in Gupta and Rawlings (2014) as an alternative to the MCMC parameter estimation.

A strategy for dealing with high-dimensional sampling problems is to combine the particle filters with the MCMC methods to obtain the sequential MCMC (SMCMC) algorithms (Septier and Peters, 2016). The MCMC methods for high-dimensional systems are compared in Septier and Peters (2016). The population MC (PMC) sampling framework to perform the Bayesian inference in high-dimensional models is developed in Koblents and Míguez (2011).

The Bayesian inference via the MC sampling utilizing the stochastic gradient descent is studied in Wang et al. (2010). The parameter likelihoods are calculated by combining the MC global sampling with the locally optimum gradient methods in Kimura et al. (2015). The nested Bayesian sampling is used in Pullen and Morris (2014) to compute the marginal likelihoods, and to compare or rank several competing models. The MCMC sampling for the mixed-effects SDE models is considered in Whitaker et al. (2017). In order to overcome the ill-conditioned least squares (LS) data fitting and the associated numerical instability problems, the bootstrapped MC procedure based on the diffusion and the LNA was proposed in Lindera and Rempala (2015).

The sequential MC (SMC) methods represent the posterior distribution by a set of samples referred to as particles (Gordon et al., 1993; Doucet et al., 2001; Tanevski et al., 2010; Yang et al., 2014), so these methods are also known as particle filters (Gordon et al., 1993; Doucet et al., 2001; Lillacci and Khammash, 2012; Golightly et al., 2015). The particle filters assume specific types of random processes to identify the posterior while bounding the computational complexity for the models with large number of parameters (Mikelson and Khammash, 2016). The particle filters are shown to be more robust than the LS data fitting, if the data statistics are exploited (Lillacci and Khammash, 2012). The SMC methods for the joint estimation of states and parameters are developed in Nemeth

et al. (2014). The degeneracy phenomenon commonly occurring in particle filters can be mitigated by more efficient sampling strategies (Golightly and Kypraios, 2017). A parallelization of the SMC computations is devised in Mihaylova et al. (2012). More efficient generation and processing of particles to improve the computational efficiency of particle filters is investigated in Golightly et al. (2019). The computationally efficient particle MCMC (pMCMC) method is devised in Koblents and Míguez (2014) and Koblents et al. (2019). The pMCMC method can be combined with the diffusion approximation (Golightly and Wilkinson, 2011), and further refined to improve its scalability (Golightly and Kypraios, 2017). The proposal distribution for the Bayesian analysis is obtained using the pMCMC sampling in Sherlock et al. (2014). The proposal samples for calculating the marginal likelihoods are obtained for the CLE and the LNA approximations in Golightly et al. (2015).

4.3. Other Statistical Methods

The key assumption for using the standard Kalman filter is the linearity of measurements. The Kalman filter is used with the CME approximation and the noise covariance estimation in Dey et al. (2018) while allowing for the dependency of the noise statistics on the states and parameter values. The Kalman filter is used to obtain the initial guess of the parameter values for the subsequent parameter estimation by data fitting in Lillacci and Khammash (2010). The Kalman filter can be merged with the particle filters to perform the inferences in stochastic (Vrettas et al., 2011) as well as deterministic systems (Arnold et al., 2014). The Kalman filter for the time integrated observations is assumed in Folia and Rattray (2018).

Since the BRNs are generally highly non-linear, the extended and unscented Kalman filters (EKFs and UKFs) must be assumed (Baker et al., 2011). The EKF was modified for stiff ODEs in Kulikov and Kulikova (2015a) and Kulikov and Kulikova (2017). The joint estimation of parameters and states by the EKF is investigated in Sun et al. (2008) and Ji and Brown (2009). The EKF is combined with the moment closure methods in Ruess et al. (2011), and it is modified for the parameter estimation in the S-system models in Meskin et al. (2011). A hybrid method combining the EKF and the particle swarm optimization (PSO) for the joint estimation of parameters and states is developed in Zeng et al. (2012). A modified EKF to penalize the modeling uncertainty due to linearization errors is proposed in Xiong and Zhou (2013) which improves the estimation accuracy. The square-root UKF achieves good numerical stability, and it can also assume the state constraints (Baker et al., 2013, 2015). For infrequent sampling and sparse observations, the UKF and the cubature Kalman filter outperform the EKF (Kulikov and Kulikova, 2015b, 2017).

The classical bootstrapping with data replication and resampling to enable the repeated estimations is described in Vanlier et al. (2013). The bootstrapping can be also used to obtain the confidence intervals of the parameter estimates (Joshia et al., 2006; Srivastavaa and Rawlingsb, 2014), and to improve the computational efficiency in recomputed model trajectories (Lindera and Rempala, 2015). The bootstrap filter can outperform the EKF (Gordon et al., 1993).

There are also many other less commonly used inference strategies which have not been mentioned so far. For instance, the Gaussian smoothing to compensate for the missing and noisy data is used in Sun et al. (2012). The parameter estimation assuming a non-linear ODE model combined with the data smoothing was investigated in J. O. Ramsay and Cao (2007). The inference of the state distribution via the optimized histograms and statistical fitting is performed in Atitey et al. (2018b). A formal verification and the sequential probability ratio test for the parameter estimation are considered in Hussain (2016). The moment closure modeling is combined with stochastic simulations for the parameter estimation in Bogomolov et al. (2015). A generalized method of moments incorporating the empirical sample moments is performed in Kügler (2012); Lück and Wolf (2016) whereas the moment based methods for the parameter estimation and the optimum experiment design are considered in Ruess and Lygeros (2015). The expectation propagation (EP) for the approximate Bayesian inference is studied in Cseke et al. (2016). The Lyapunov exponent can be used to infer the level of predictability of the dynamic systems including BRNs (Barnes et al., 2011; McGoff et al., 2015).

4.4. Model Fitting Methods

The parameter estimation by fitting the measured data appears to be by far the most commonly used method in literature. The main reason is that, unlike other estimation strategies, the data fitting problem is relatively easy to formulate with minimum knowledge and assumptions. It is possible to consider multiple fitness functions. Various continuous and discrete fitness functions are explored in Deng and Tian (2014). The fitness function can be derived from the likelihood function (Rodriguez-Fernandez et al., 2006a), or the approximated likelihood function (Srivastavaa and Rawlingsb, 2014).

Even though the derivative free methods are easier to implement, the gradient based methods have faster albeit only local convergence. For instance, the gradient based optimization with sensitivity analysis assuming finite differences is investigated in Loos et al. (2016). The derivative free methods are necessary for the combinatorial and the integer constrained problems (Cedersund et al., 2016; Gábor et al., 2017).

The challenge is to develop numerically efficient methods to solve high-dimensional problems with possibly many constraints. The observations are interpolated with the spline functions in Nim et al. (2013), so that the derivatives can be used to estimate the production and consumption of molecules in BRNs. It decomposes a high-dimensional problem into the product of low-dimensional factors. The fitness function is interpolated with the spline functions in Zhan and Yeung (2011).

The data fitting is generally more computationally demanding for stochastic than for deterministic models, but the former is more likely to find a global solution (Rodriguez-Fernandez et al., 2006b). Since many practical optimization problems are non-convex, the global optimization methods are generally preferred. They can be implemented as multi-start or multi-shooting local methods, or by selecting a subset of parameters to be estimated. The sensitivity to initial values can be reduced by tracking multiple solutions. Many of these methods can be readily

TABLE 6 | Common evolutionary algorithms for the parameter estimation in BRNs and dynamic systems.

Algorithm	Motivation and selected papers
Genetic algorithms (GAs)	Largest class of EAs, inspired by evolution and natural selection, often near optimum solution
	Matsubara et al., 2006; Tian et al., 2007; Besozzi et al., 2009; Chou and Voit, 2009; Liu et al., 2012; Sun et al., 2012
Genetic programming (GP)	Evolution of computer programs toward improving their fitness to solve a given task
	Chou and Voit, 2009; Sun et al., 2012; Nobile et al., 2013
Evolutionary programming (EP)	Parameters of computer program evolve toward improving its fitness to solve a given task
	Baker et al., 2010; Sun et al., 2012; Revell and Zuliani, 2018
Simulated annealing (SA)	Probabilistic search combining sampling with random but controlled acceptance of candidate solutions
	Ashyraliyev et al., 2009; Chou and Voit, 2009; Dai and Lai, 2010; Sun et al., 2012; Hussain et al., 2015; Cedersund et al., 2016
Differential evolution (DE)	Derivative free method, linearly combining randomly selected candidate solutions to obtain iterative improvements
	Srinivas and Rangaiah, 2007; Liu and Wang, 2009; Chong et al., 2012, 2014; Sun et al., 2012; Teijeiro et al., 2017
Scatter search (SS)	Often combined with tabu search, it is local search with temporarily accepting worse solutions and avoiding already visited regions
	Rodriguez-Fernandez et al., 2006a; Villaverde et al., 2012; Cedersund et al., 2016; Penas et al., 2017; Remlia et al., 2017
Particle swarm optimiz. (PSO)	Derivative free method, moving particles (i.e., samples or candidate solutions) toward better solution
	Besozzi et al., 2009; Abdullah et al., 2013c; Sun et al., 2014; Cazzaniga et al., 2015; Nobile et al., 2016; Tangherloni et al., 2016

parallelized to overcome the computational burden (Mancini et al., 2015; Teijeiro et al., 2017). The parallel implementations of data fitting algorithms including Spark, MapReduce, and MPI messaging are considered in Teijeiro et al. (2017). Recently, the implementations exploiting the affordable graphical processing units (GPUs) have become popular (Nobile et al., 2012). The computational complexity of global methods can be mitigated by the incremental identification strategies (Michalik et al., 2009). The global methods also require to properly set the search parameters which can be done via multiple initial exploratory runs (Penas et al., 2017). Another global search strategy assumes a model transformation followed by the non-uniform sampling (Kleinstein et al., 2006). There are also hybrid strategies switching between the global and local searches (Rodriguez-Fernandez et al., 2006a,b; Ashyraliyev et al., 2009).

The majority of data fitting methods are rooted in the simple LS regression, or assume the non-linear least squares (NLSQ) (Baker et al., 2011). The alternating regression (AR) reformulates the non-linear fitting as an iterative linear regression problem (Chou et al., 2006). The non-linear regression is converted into a non-linear programming problem which is solved by the random drift PSO in Sun et al. (2014). The asymptotic properties of the LS estimation were evaluated in Rempala (2012). The iterative linear LS for systems described by a ratio of linear functions is considered in Tian et al. (2010).

The regularization is a strategy to deal with the ill-conditioned optimization problems due to insufficient or noisy data (Gábor and Banga, 2014; Gábor et al., 2017). The regularization introduces additional constraints to penalize the complexity, or it uses prior knowledge to constrain the parameter values to trade-off the estimator bias with its variance in order to avoid the model overfitting (Liu et al., 2012; Kravaris et al., 2013; Jang et al., 2016). Alternatively, the perturbation method has been developed for fitting the data in Shiang (2009).

The evolutionary algorithms (EAs) are the most frequently used methods for solving the high-dimensional constrained optimization problems. They do not require any particular assumptions, and they are not limited by the dimensionality of the problem. The EAs adopt various heuristic strategies to find the optimum assuming the population of candidate solutions which are iteratively improved by reproduction, mutation, crossover or recombination, selection and other operations until the fitness or loss function reaches the desired value. The specific EAs commonly used in literature for the identification of BRNs and other dynamic systems are summarized in **Table 6**. Several EAs and the PSO methods are compared in Nobile et al. (2018b). Different EAs are compared with other deterministic search methods in Mendes and Kell (1998).

The cuckoo search utilizes random sub-populations which can be discarded to improve the solution (Rakhshania et al., 2016). The optimization programs include non-linear simplex method (Cazzaniga et al., 2015), non-linear programming (NLP) (Moles et al., 2003; Zhan and Yeung, 2011; Sun et al., 2012; Rodriguez-Fernandez et al., 2013), semi-definite programming (Kuepfer et al., 2007; Rumschinski et al., 2010), and quadratic programming (Gupta, 2013). The Nelder-Mead method (also known as the downhill simplex method) maintains a simplex of the test points which evolve until the data fit is found (Abdullah et al., 2013a). The quantifier elimination (QE) is used to simplify the constrained optimization problems (Anai et al., 2006). Other examples of the nature inspired algorithms include the firefly algorithm (FA) (Abdullah et al., 2013a,b) and the artificial bee colony (ABC) algorithm (Chong et al., 2014). Neural networks are becoming popular especially due to multi-layer deep learning methods. Other tasks encountered in traditional neural networks involve training, overfitting, smoothing, and the mean value approximations (Matsubara et al., 2006; Chou and Voit, 2009; Ali et al., 2015; Berrones et al., 2016). The parallel implementation of the scatter search for large-scale systems is devised in Villaverde et al. (2012) and Penas et al. (2017).

The benefits of individual optimization methods can be utilized by adaptively combining different algorithms. For

TABLE 7 | The coverage of the parameter estimation methods for BRNs.

	Tasks				Measures					Bayesian methods				Monte Carlo			Kalman filter			Model fitting							XLR
	Identifi., observab., reachability	Optimum experiment design	Bifurcation analysis	Inference, identification	Sensitivity analysis	Confidence/credible intervals	Akaike/Fisher/mutual info	Entropy	Sum of squared errors	MAP, ML, likelihood	Approximate Bayesian comput	Expectation-maximization	Variational Bayesian inference	MCMC	Metropol./import. sampling	Sequential MC, particle filters	Kalman filter	Extended Kalman filter	Unscented Kalman filter	LS and regression	Genetic algorithms	Optimization programming	Simulated annealing	Differential evolution	Scatter, tabu, cuckoo search	Particle swarm optimization	Other algorithms	Mach./deep/transf. learning
# papers	148	62	9	297	83	63	70	52	21	238	31	59	34	84	80	79	89	54	30	112	82	196	99	87	117	45	73	46

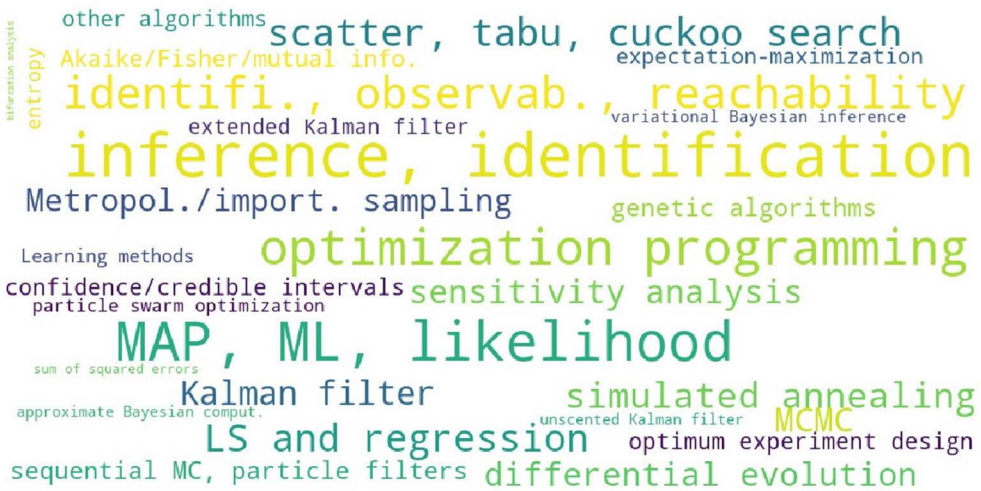


FIGURE 3 | A word cloud visualizing the levels of interest in different parameter estimation methods and tasks for models of BRNs.

instance, the DE is combined with the tabu search in Srinath and Gunawan (2010), and another hybrid DE method is considered in Liu and Wang (2008b). The genetic programming and the PSO are combined in Nobile et al. (2013), the multi-swarm PSO is considered in Nobile et al. (2012), and the fuzzy logic based PSO is developed in Nobile et al. (2015), Nobile et al. (2016), and Nobile et al. (2018a). The regularization, pruning and the continuous genetic algorithm (CGA) are combined in Liu et al. (2012).

Machine learning (MLR) methods can be very effective provided that there is enough training data drawn from some fixed distribution (Pan and Yang, 2010). If there are not enough labeled data, or the generating distribution changes in time, it may be better to employ transfer learning (TLR) methods which exploit data from multiple domains (Pan and Yang, 2010; Weiss et al., 2016; Azab et al., 2018). A primer on the MLR and the deep learning (DLR) methods for biological networks is provided in Camacho et al. (2018).

The survey of 5 estimation tasks and 23 estimation methods for BRNs identified in the references listed at the end of this paper is provided in **Table S2**. This table is summarized in **Table 7** for convenience, and the corresponding word cloud is shown in **Figure 3**. Other tasks related to the parameter estimation which are commonly used in literature are the model identifiability, the parameter observability, and the reachability analysis. The information theoretic measures are assumed

relatively often as an alternative to the probabilistic measures to define the rigorous inference problems. The parameter identification by model fitting appears to be the most common strategy in literature. The Bayesian analysis which accounts for the prior distribution of parameters is often performed numerically by adopting the MCMC and other statistical i sampling methods.

In order to visualize a timeline of interest in different parameter estimation methods, **Table 8** contains the numbers of cited papers concerning the specific estimation methods and tasks in given years. As for the methods in **Table 3**, we can observe that the general interest appears to have peaked in 2014, although the considerable interest has remained strong over the past decade. This indicates that the parameter estimation strategies are closely related to the modeling strategies as discussed previously.

5. CHOICES OF MODELS AND METHODS FOR INFERENCES IN BRNS

We now evaluate what BRN models are preferred with the different parameter estimation strategies, and also explore what parameter estimation methods are assumed in different parameter estimation tasks. The models and the estimation tasks and methods are the same as those considered in **Tables 2**, **7**, respectively.

Table 9 shows the number of papers concerning given BRN models and given estimation strategies. The paper counts were adjusted to exclude papers which were deemed to only marginally consider a given combination of the BRN model and the estimation task or method. In particular, the papers containing <5 occurrences of the search keywords for either a given model, task or method were excluded. We can observe that the parameter inference tasks have been considered for all the BRN models, however, some models have been investigated much more than the others. The most popular models for the parameter inferences and other related tasks are the models involving differential equations, Markov processes, and state space representations. The second most popular group of models considered for the parameter estimation include the S-system and polynomial models, and the moment closure and the LNA models.

The sensitivity analysis using the information theoretic measures and evaluation of the confidence and credible intervals have been considered for most BRN models. The sensitivity analysis has somewhat similar use of models as the parameter inference, except the level of interest in the former is about ten times smaller. Moreover, the sensitivity analysis is often combined with the bifurcation analysis, so the latter may not be referred to explicitly in many papers. The optimum experiment design has been assumed for several models, but there seems to be no clear model preference. The sum of squares measure is likely quite underestimated in **Table 9**, since it is often assumed without being explicitly referred to.

The probabilistic MAP and ML measures have been assumed for all model types. In many cases, the corresponding inference tasks involve the prior and posterior distributions and probabilities, and the parameter likelihoods. The variational Bayesian and the ABC methods are mostly used with the Markov processes, since this is where they were originally developed for whereas the Markov processes can be derived from differential equations. The EM method is mostly used with the differential equations. The MC based sampling methods including particle filters are important for practical implementation of the Bayesian inference strategies. However, these methods seem to be rarely used with less popular BRN models. Similar comments can be made about the Kalman filtering, the LS regression, and most of the data fitting methods considered. The PSO method has been mainly considered with the models involving differential equations, and to some extent also with several other models. There are several BRN models which are not assumed with other inference algorithms, such as neural networks.

The statistical learning methods including MLR, DLR and TLR are still used sporadically compared to the other methods discussed so far. Consequently, it is still difficult to identify which BRN models in literature are preferred for statistical learning. The statistical learning requires enough training data as well as some level of time invariance in order to find generalized descriptions of systems, and to make predictions from the data. However, as the interest in applications of the MLR techniques continues to grow, and the efficiency of learning from data improves, it will also affect suitability of the MLR techniques for use with the different BRN models.

Another interesting viewpoint is to evaluate what inference methods are used for different inference tasks. The numbers of papers for given combinations of the inference tasks and the inference methods are provided in **Table 10**. With one exception, there is at least one paper for each such combination, however, the level of interest varies considerably. In particular, the largest number of papers for all the inference tasks considered assume the Bayesian analysis and the methods for the model fitting to data. On the other hand, the sum of squared errors, the UKF, and the PSO methods are generally the least assumed. As discussed previously, the sum of squared errors is used often, but rarely mentioned explicitly whereas the UKF and the PSO methods are usually rather difficult to implement.

Assuming **Table 10**, we can compare the levels of interest for two or more methods and the given inference task. For example, the EM and the MCMC methods are used equally often for the sensitivity analysis whereas the MCMC method is preferred over the EM method for the identifiability task. The LS and the regression methods seem to be always preferred over Kalman filtering due to its implementation complexity. Interestingly, the MLR methods appear to be considered more often than the ABC, the variational Bayesian inference, the UKF, and the PSO methods, but comparably often to the EKF.

5.1. Future Research Directions

Tables 9, **10** together with the data in **Tables 2**, **7** can be used as guidelines to define new research problems which have not been sufficiently investigated in literature. We can separate the models, tasks and methods into the groups according to their levels of interest. Due to sparsity of data in **Table 9**, it

TABLE 8 | The number of papers concerning the estimation tasks and methods for BRNs in given years.

	Tasks				**Measures**					**Bayesian methods**				**Monte Carlo**			**Kalman filter**			**Model fitting**								**XLR**
year	Identifi., observab., reachability	Optimum experiment design	Bifurcation analysis	Inference, identification	Sensitivity analysis	Confidence/credible intervals	Akaike/Fisher/mutual info.	Entropy	Sum of squared errors	MAP, ML, likelihood	Approximate Bayesian comput.	Expectation-maximization	Variational Bayesian inference	MCMC	Metropol./import. sampling	Sequential MC, particle filters	Kalman filter	Extended Kalman filter	Unscented Kalman filter	LS and regression	Genetic algorithms	Optimization programming	Simulated annealing	Differential evolution	Scatter, tabu, cuckoo search	Particle swarm optimization	Other algorithms	Mach./deep/transf. learning
2005	3	2	.	7	1	1	1	.	.	3	.	.	.	2	2	1	2	.	.	1	1	3	1	1	.	.	.	.
2006	6	1	.	10	1	4	2	.	2	7	.	1	1	.	.	.	.	.	.	4	3	4	4	6	1	.	3	1
2007	2	2	1	8	1	2	1	.	2	5	.	.	.	1	1	2	2	1	1	4	4	5	4	3	1	1	2	2
2008	3	3	.	9	1	1	1	.	1	6	.	2	.	1	1	1	2	1	1	5	3	7	3	5	4	.	3	1
2009	8	4	1	13	4	4	4	2	1	11	1	.	2	3	3	2	4	1	2	7	8	9	7	7	6	3	4	1
2010	13	5	.	22	8	4	8	3	2	18	3	4	2	5	8	6	5	3	1	7	7	16	10	8	6	2	4	3
2011	9	5	.	18	6	3	2	1	.	15	1	4	2	6	7	4	4	4	2	4	4	11	6	4	6	2	2	.
2012	14	7	3	27	10	5	10	6	.	21	2	9	3	8	12	10	6	3	2	11	9	17	10	10	13	6	10	4
2013	19	6	2	30	11	6	11	8	3	28	3	4	4	12	6	10	14	11	2	14	12	21	9	7	15	10	9	8
2014	27	12	1	45	13	17	12	10	4	40	6	14	10	20	14	19	19	12	8	21	11	27	15	9	18	5	10	8
2015	16	3	.	28	8	4	3	4	1	22	5	8	1	9	7	10	13	8	7	11	5	20	7	5	8	4	6	3
2016	12	2	1	28	9	6	6	5	3	24	3	5	1	7	6	4	8	4	1	11	6	18	10	9	14	6	7	7
2017	10	7	.	24	7	6	5	6	2	16	3	4	6	5	6	3	5	2	1	7	3	20	7	8	16	4	7	3
2018	7	2	.	21	3	1	4	7	.	17	3	3	1	4	5	6	3	2	2	4	4	14	4	4	7	2	6	4

TABLE 9 | The adjusted number of papers concerning given estimation methods and given BRN models.

		Tasks				Measures						Bayesian methods				Monte Carlo			Kalman filter	Model fitting												XLR
		Identifi., observab., reachability	Optimum experiment design	Bifurcation analysis	Inference, identification	Sensitivity analysis	Confidence/credible intervals	Akaike/Fisher/mutual info.	Entropy	Sum of squared errors	MAP, ML, likelihood	Approximate Bayesian comput.	Expectation-maximization	Variational Bayesian inference	MCMC	Metropol./import. sampling	Sequential MC, particle filters	Kalman filter	Extended Kalman filter	Unscented Kalman filter	LS and regression	Genetic algorithms	Optimization programming	Simulated annealing	Differential evolution	Scatter, tabu, cuckoo search	Particle swarm optimization	Other algorithms	Mach./deep/transf. learning
Physical laws	Kinetic rate laws	3	2	.	10	3	1	.	2	.	6	1	.	1	1	.	2	2	2	2	3	2	4	1	3	3	1	2	.
	Mass action kinetics	7	2	.	19	2	1	.	2	1	11	.	3	2	4	1	.	.	.	.	4	2	4	2	3	2	1	2	.
	Mechanistic models	5	.	.	15	2	5	1	1	1	12	3	3	.	3	1	3	1	1	.	2	1	4	2	2	2	2	2	1
Random proc.	Markov process	22	3	1	100	10	11	8	9	1	86	15	8	8	50	30	27	11	4	3	17	3	12	5	3	7	.	5	4
	Poisson process	6	2	1	22	4	2	3	6	1	17	2	3	4	12	9	4	2	.	.	4	2	4	.	1	3	.	1	.
	Birth-death process	3	1	.	9	1	1	.	1	.	8	.	1	2	3	2	1	.	.	.	.	.	.	.	.	1	.	.	.
	Telegraph process	.	.	.	2	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.
	state space repres.	16	5	1	54	7	4	5	6	1	45	5	5	3	15	13	14	16	8	7	7	2	8	3	1	5	1	3	2
	ODE,PDE,SDE,DDE	43	8	1	133	16	15	12	8	3	90	8	14	6	32	16	20	19	12	9	19	9	26	12	15	14	9	9	3
Mathematical models	Rational model	4	.	.	6	2	2	1	.	.	5	2	1	.	2	2	4	1	1	1	2	1	3	3	1	1	.	1	2
	Diff. algebraic eqns.	4	1	.	6	1	1	.	.	.	4	1	.	.	.	1	.	1	.	.	3	3	3	2	2	.	1	.	.
	Tensor representation	3	2	.	4	2	1	1	1	.	3	.	.	.	.	.	.	1	1	1	1	.	.	1	.	.	.	1	.
	S-system model	4	2	1	25	3	2	1	2	.	9	.	2	2	2	1	.	3	3	2	11	5	6	3	7	.	2	2	.
	Polynomial model	10	3	1	26	5	4	3	4	1	13	1	2	2	5	4	1	2	1	2	9	2	11	.	2	1	.	2	.
	Manifold map	3	.	.	7	2	1	.	1	.	6	3	2	.	3	2	4	3	2	2	4	.	.	1	.	1	.	3	2
Interaction models	Petri nets	1	.	1	3	2	.	1	2	.	2	.	1	.	1	1	.	.	.	.	1	1	1	.	.	.	.	1	.
	Boolean networks	2	.	1	2	1	1	2	1	.	2	.	1	.	1	1	1	.	.	.	2	1	2	.	1	.	.	.	1
	Neural networks	3	.	.	9	3	1	1	2	.	5	1	1	.	1	1	1	3	3	3	5	2	2	2	1	1	1	9	3
	Agent based models	2	.	1	9	2	1	1	2	.	7	1	2	.	3	3	5	2	.	.	1	2	3	2	.	2	.	1	.
CME based models	Langevin equation	4	1	.	17	4	2	2	3	1	15	.	.	3	9	8	4	3	1	.	.	.	2	1	.	3	.	1	.
	Fokker-Planck equation	5	2	.	11	1	2	2	2	1	9	.	.	4	4	2	1	1	.	.	.	.	1	.	.	3	.	.	.
	Reaction rate equation	3	1	.	3	1	1	1	.	.	3	.	.	.	2	1	.	1	.	.	.	.	1	.	.	2	.	.	.
	Moment closure	8	.	.	24	4	1	3	5	2	17	1	1	4	5	2	3	2	1	.	1	.	2	.	.	5	.	1	1
	Linear noise approx.	10	1	.	29	6	3	4	4	2	25	2	2	1	12	5	6	2	.	.	1	.	.	.	1	5	.	1	.
	System size expansion	2	1	.	4	1	1	.	1	2	4	.	1	1	2	.	.	.	.	.	.	.	.	.	.	1	.	.	.

TABLE 10 | The number of papers concerning given tasks and given estimation methods for BRNs.

	Measures			Bayesian methods		Monte Carlo			Kalman filter			Model fitting								XLR
	Confidence/credible intervals	Akaike/Fisher/mutual info.	Entropy	Sum of squared errors	MAP, ML, likelihood	Approximate Bayesian comput.	Expectation-maximization	Variational Bayesian inference	MCMC	Metropol./import. sampling	Sequential MC, particle filters	Kalman filter	Extended Kalman filter	Unscented Kalman filter	LS and regression	Genetic algorithms	Optimization programming	Simulated annealing	Differential evolution	Searches	Particle swarm optimiz.	Other algorithms	Mach./deep/transf. learning
Identifi., observ., reach.	44	54	28	10	131	20	25	21	45	37	43	49	27	15	71	42	96	56	45	60	19	30	26
Bifurcation analysis	17	15	15	9	55	10	17	11	25	24	23	16	8	4	28	22	49	31	16	25	10	15	9
Optimum experiment	4	7	3	2	9	0	4	2	2	3	2	3	1	1	5	4	7	4	5	4	2	5	3
Inference, identif.	65	70	52	21	236	30	59	34	83	79	78	89	54	30	113	82	197	99	87	116	45	73	46
Sensitivity analysis	25	36	19	6	72	13	24	15	22	21	27	28	17	12	44	30	65	37	31	35	15	19	16

is easier to enumerate the problems which have already been well-investigated in the literature. Such cases are highlighted in **Table 9**, and they include:

- The identification and inference tasks with the Markov processes, state space representations, differential equations, polynomial function, S-system, Langevin and Fokker-Planck equations, and the CME approximation models;
- Most of the inference methods with the Markov processes, state space representations, and the differential equation models;
- Some inference methods assuming the Poisson process, S-system, polynomial function, Langevin equation, and the CME approximation models;
- The Bayesian methods with the MAP and ML inferences with most of the models considered;
- The LS regression and the optimization programming mainly with the Markov processes, state space representations, differential equations, S-system and the polynomial models; and
- The search methods with the Markov processes, state space representations, differential equations, and the CME approximation models.

The bifurcation analysis appears to be the least considered task for all models. However, in many papers, the bifurcation analysis may not be referred to explicitly as it is performed as part of the sensitivity analysis. Similar comments can be made about the sum of squared errors. From **Table 9**, we observe that also machine learning methods have been considered sporadically and only for some BRN models to solve the inference problems. Comparing machine learning methods with the conventional methods of statistical inference may be one of the most interesting research avenues in near future. It is likely that machine learning is more beneficial for some models, depending on the availability of observations and training data. In addition, we can observe from **Table 10** that the optimum experiment design did not receive as much attention in literature as other inference tasks.

There are likely other research opportunities which are not immediately apparent from the tables in previous sections. For instance, the minimum mean square error (MMSE) estimator is only discussed in the reference (Koeppl et al., 2012). Since the estimation errors may have different distributions depending on the BRN model considered, the generalized linear regression (GLR) can be assumed as a simple to implement, universal and yet powerful statistical learning technique. The GLR method has not been investigated for the inferences in BRNs. It is also useful to estimate other quantities in addition to inferring the parameter values. For example, the distributions of species counts are estimated in Atitey et al. (2018b). Knowledge of the parameter distributions greatly affects the available choices of estimators and their performance. Another unexplored strategy is the compressive sensing (CS) which exploits the sparsity in parameter spaces. Among machine learning methods, the transfer learning has not been used for inferences in BRNs in order to exploit the increasing production of omics data (Weiss et al., 2016).

Furthermore, the vast majority of inference problems in literature assume the well-stirred models of BRNs with the reactions dependent solely on the species concentrations, but not on the species spatial distributions. Assuming the spatially resolved models of BRNs with the diffusion and other phenomena of the molecular transport through complex fluids is much more realistic. Such models are usually described by the RDME (Lötstedt, 2018). Moreover, in many BRNs, the reaction rates are time varying. The inferences of time varying parameters in BRN models have not been explicitly considered in literature.

Most inference problems in literature assume simple models of measurements, such as obtaining the noisy concentrations of species at discrete time instances. In order to increase the sensitivity of measurements, the observations are often accumulated in time (Folia and Rattray, 2018). The transformations, such as the time integration of measurements must be incorporated into the BRN models when devising the interference strategies. Since the measurements may affect the biological processes, the number and duration of the measurements should be minimized in space and in time. In addition, the measurement noise is often (but not always) assumed to be independent of the species concentrations and Gaussian distributed. In realistic *in vivo* and *in vitro* experiments, the measurement noises are correlated in time and with other measurements, and also dependent on the reaction rates and the species concentrations. It would be very useful to report the statistical properties of measurements from the different laboratory experiments. Having such statistical description of measurements can considerably improve the efficiency and accuracy of the inference methods in BRNs.

More generally, the performance of various inference strategies is greatly dependent on the structure, parameter values and the initial state of the BRN considered. These aspects were considered mostly to optimize the data fitting methods, but much less for the other inference methods. There is a trade-off in mechanistically employing the universal inference methods, and adopting these methods to specific scenarios of the BRNs. The latter approach may improve the performance and efficiency of the parameter inference at the cost of increased implementation complexity. More research is needed to jointly explore the model simplification strategies and the parameter estimation strategies as in Eghtesadi and Mcauley (2014). However, it is always important to test and validate all the inference algorithms devised. In some papers, the inference algorithms are tested on multiple data sets, but a general methodology for testing and validating the inference algorithms for BRNs have not been presented in literature.

Many papers on the inferences in BRNs are concerned with the implementation aspects rather than the concepts. It would be useful to separate the inference concepts and strategies form their implementation. For example, the Bayesian inference can be implemented using the stochastic sampling, the ABC, the variational inference, the EM and several other methods.

Finally, let's not forget that the ultimate goal of performing the statistical inferences in BRNs is to improve our understanding of the *in vivo* and *in vitro* biological systems and phenomena. It is primarily dependent on having the sufficiently accurate models of these systems including knowing the values of their parameters. As the experimental techniques improve, the new data from the experiments will likely stimulate the developments of new biological models, and thus, there will also be the need for new inference methods and strategies.

6. CONCLUSIONS

The aim of this review paper was to explore how various inference tasks and methods are used with different models of BRNs. The key concepts of modeling and the parameter inferences for BRNs were discussed. The dependency between tasks, methods and models were captured in tables containing the paper counts. More detailed information is provided in **Supplementary Tables** including the links for selected papers to their citations in Google Scholar.

The common models and inference tasks and methods for BRNs were identified by text mining the cited references. The text mining was partly automated using text processing scripts. Such automation is indispensable when dealing with a large number of references as is the case in this paper. For convenience, the identified models and methods were presented under several loosely defined categories. The most common models of BRNs in literature are the mass action kinetics, Markov processes, state space representations, and differential equations. Somewhat less common, but still popular models include the kinetic rate law, mechanistic models, Poisson processes, polynomial and rational functions, the S-system model, the Langevin equation, and the CME based approximation models.

Several previously published review papers concerning the inferences in BRNs were listed. The relevant graduate research theses from the past decade were also outlined, since they tend to contain comprehensive literature surveys and tutorial style explanations. We observed that the most common inference tasks are concerned with the model identifiability, the parameter inference and the sensitivity analysis. The most common inference methods are the Bayesian analysis using the MAP and ML estimators, the MC sampling techniques, the LS regression, and the evolutionary algorithms for data fitting including the optimization programming, the simulated annealing, and the scatter and other searches.

In the last part of the paper, the levels of interest in different inference tasks and methods for given BRN models were assessed. This allowed us to identify the inference problems for BRNs which were less explored in the literature previously. Our study revealed that the interest in the inference problems in BRNs peaked in 2014. This may indicate that development of the traditional statistical methods has saturated, and the current focus is more on their efficient implementation, especially to process the massive amounts of data. The new developments will likely be driven by the machine learning methods and the continuing progress experimental techniques. The results presented in this review can be used to develop a coherent theory comprising the models and methods for the statistical inferences in BRNs.

AUTHOR CONTRIBUTIONS

All authors: substantial contributions to the conception and design of the work. PL: drafting the work. All authors: revising the work critically for important intellectual content, final approval of the version to be published, and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

KA was supported by the Zienkiewicz Scholarship award from Swansea University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00549/full#supplementary-material

REFERENCES

Abdullah, A., Deris, S., Anwar, S., and Arjunan, S. N. (2013a). An evolutionary firefly algorithm for the estimation of nonlinear biological model parameters. *PLoS ONE* 8:e56310. doi: 10.1371/journal.pone.0056310

Abdullah, A., Deris, S., Mohamad, M. S., and Anwar, S. (2013b). An improved swarm optimization for parametera estimation and biological model selection. *PLoS ONE* 8:e61258. doi: 10.1371/journal.pone.0061258

Abdullah, A., Deris, S., Mohamad, M. S., and Hashim, S. Z. M. (2013c). A new particle swarm evolutionary optimization for parameter estimation of biological models. *IJCISIM* 5, 571–580.

Ale, A., Kirk, P., and Stumpf, M. P. (2013). A general moment expansion method for stochastic kinetic models. *Chem. Phys.* 138, 1–12. doi: 10.1063/1.4802475

Ali, J. M., Hoang, N. H., Hussain, M. A., and Dochain, D. (2015). Review and classification of recent observers applied in chemical process systems. *Comput. Chem. Eng.* 76, 27–41. doi: 10.1016/j.compchemeng.2015.01.019

Amrein, M., and Künsch, H. R. (2012). Rate estimation in partially observed Markov jump processes with measurement errors. *Stat. Comput.* 22, 513–526. doi: 10.1007/s11222-011-9244-1

Anai, H., Orii, S., and Horimoto, K. (2006). Symbolic-numeric estimation of parameters in biochemical models by quantifier elimination. *Bioinform. Comput. Biol.* 4, 1097–1117. doi: 10.1142/S0219720006002351

Andreychenko, A., Mikeev, L., Spieler, D., and Wolf, V. (2011). *Parameter Identification For Markov Models Of Biochemical Reactions*, chapter CAV, Vol. 6806. Berlin; Heidelberg: Springer, 83–98.

Andreychenko, A., Mikeev, L., Spieler, D., and Wolf, V. (2012). Approximate maximum likelihood estimation for stochastic chemical kinetics. *EURASIP J. Bioinf. Sys. Biol.* 2012:9. doi: 10.1186/1687-4153-2012-9

Andreychenko, A., Mikeev, L., and Wolf, V. (2015). Model reconstruction for moment-based stochastic chemical kinetics. *ACM Trans. Model. Comput. Simul.* 25, 12:1–12:19. doi: 10.1145/2699712

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. B* 72, 269–342. doi: 10.1111/j.1467-9868.2009.00736.x

Angius, A., and Horváth, A. (2011). The Monte Carlo EM method for the parameter estimation of biological models. *El. Notes Theor. Comp. Sci.* 275, 23–36. doi: 10.1016/j.entcs.2011.09.003

Arnold, A., Calvetti, D., and Somersalo, E. (2014). Parameter estimation for stiff deterministic dynamical systems via ensemble Kalman filter. *Inv. Prob.* 30:105008. doi: 10.1088/0266-5611/30/10/105008

Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2009). Systems biology: parameter estimation for biochemical models. *FEBS J.* 276, 886–902. doi: 10.1111/j.1742-4658.2008.06844.x

Atitey, K., Loskot, P., and Rees, P. (2018a). Determining transcription rates yielding steady state production of mRNA in the lac genetic switch of *Escherichia coli*. *Comput. Biol.* 25, 1023–1039. doi: 10.1089/cmb.2018.0055

Atitey, K., Loskot, P., and Rees, P. (2018b). Inferring distributions from observed mRNA and protein copy counts in genetic circuits. *Biomed. Phy. Eng. Express* 5:015022. doi: 10.1088/2057-1976/aaef5c

Atitey, K., Loskot, P., and Rees, P. (2019). Elucidating effects of reaction rates on dynamics of the lac circuit in *Escherichia coli*. *Biosyst.* 175, 1–10. doi: 10.1016/j.biosystems.2018.11.003

Azab, A. M., Toth, J., Mihaylova, L. S., and Arvaneh, M. (2018). "A review on transfer learning approaches in brain-computer interface," in *Signal Processing and Machine Learning for Brain-Machine Interfaces*, eds T. Tanaka and M. Arvaneh (London, UK: IET), 81–98.

Babtie, A. C., and Stumpf, M. P. H. (2017). How to deal with parameters for whole-cell modelling. *J. R. Soc. Interface* 14:237. doi: 10.1098/rsif.2017.0237

Baker, S. M., Poskar, C. H., and Junker, B. H. (2011). Unscented Kalman filter with parameter identifiability analysis for the estimation of multiple parameters in kinetic models. *EURASIP Bioinf. Sys. Biol.* 2011:7. doi: 10.1186/1687-4153-2011-7

Baker, S. M., Poskar, C. H., Schreiber, F., and Junker, B. H. (2015). A unified framework for estimating parameters of kinetic biological models. *BMC Bioinformatics* 16:9. doi: 10.1186/s12859-015-0500-9

Baker, S. M., Schallau, K., and Junker, B. H. (2010). 133, Comparison of different algorithms for simultaneous estimation of multiple parameters in kinetic metabolic models. *Integr. Bioinf.* 7:133. doi: 10.2390/biecoll-jib-2010-133

Banga, J. R., and Balsa-Canto, E. (2008). Parameter estimation and optimal experimental design. *Essays Biochem.* 45, 195–210. doi: 10.1042/bse0450195

Barnes, C. P., Silk, D., and Stumpf, M. P. (2011). Bayesian design strategies for synthetic biology. *Interface Focus* 1, 895–908. doi: 10.1098/rsfs.2011.0056

Bayer, C., Moraes, A., Tempone, R., and Vilanova, P. (2016). An efficient forward-reverse expectation-maximization algorithm for statistical inference in stochastic reaction networks. *Stoch. Analysis Appl.* 34, 193–231. doi: 10.1080/07362994.2015.1116396

Berrones, A., Jiménez, E., Alcorta-García, M. A., Almaguer, F.-J., and na, B. P. (2016). Parameter inference of general nonlinear dynamical models of gene regulatory networks from small and noisy time series. *Neurocomputing* 175, 555–563. doi: 10.1016/j.neucom.2015.10.095

Besozzi, D., Cazzaniga, P., Pescini, G. M. D., and Vanneschi, L. (2009). "A comparison of genetic algorithms and particle swarm optimization for parameter estimation in stochastic biochemical systems," in *EvoBIO* (Berlin: Springer), 116–127.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Bogomolov, S., Henzinger, T. A., Podelski, A., Ruess, J., and Schilling, C. (2015). "Adaptive moment closure for parameter inference of biochemical reaction networks," in *CMSB 2015* (Cham: Springer), 77–89.

Bouraoui, I., Farza, M., Ménard, T., Abdennour, R. B., M'Saad, M., and Mosrati, H. (2015). Observer design for a class of uncertain nonlinear systems with sampled outputs application to the estimation of kinetic rates in bioreactors. *Automatica* 55, 78–87. doi: 10.1016/j.automatica.2015.02.036

Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.* 18, 125–135. doi: 10.1007/s11222-007-9043-x

Brim, L., Češka, M., Dražan, S., and Šafránek, D. (2013). On robustness analysis of stochastic biochemical systems by probabilistic model checking. arXiv:1310.4734v1 [cs.NA].

Bronstein, L., and Koeppl, H. (2018). A variational approach to moment-closure approximations for the kinetics of biomolecular reaction networks. *J. Chem. Phys.* 148:014105. doi: 10.1063/1.5003892

Bronstein, L., Zechner, C., and Koeppl, H. (2015). Bayesian inference of reaction kinetics from single-cell recordings across a heterogeneous cell population. *Methods* 85, 22–35. doi: 10.1016/j.ymeth.2015.05.012

Busetto, A. G., and Buhmann, J. M. (2009). "Stable Bayesian parameter estimation for biological dynamical systems," in *ICCSE* (Vancouver, BC), 148–157.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173, 1–12. doi: 10.1016/j.cell.2018.05.015

Carmi, A., Mihaylova, L., Gning, A., Gurfil, P., and Godsill, S. (2013). "Markov chain Monte Carlo based autonomous tracking and causality reasoning," in *Advances in Intelligent Signal Processing and Data Mining* (Berlin; Heidelberg: Springer), 7–53.

Cazzaniga, P., Nobile, M. S., and Besozzi, D. (2015). "The impact of particles initialization in PSO: parameter estimation as a case in point," in *CIBCB* (Niagara Falls, ON), 1–8.

Cedersund, G., Samuelsson, O., Ball, G., Tegnér, J., and Gomez-Cabrero, D. (2016). "Optimization in biology parameter estimation and the associated optimization problem," in *Uncertainty in Biology* (Cham: Springer), 177–197.

Češka, M., Dannenberg, F., Paoletti, N., Kwiatkowska, M., and Brim, L. (2017). Precise parameter synthesis for stochastic biochemical systems. *Acta Informatica* 54, 589–623. doi: 10.1007/s00236-016-0265-2

Ceska, M., Safránek, D., Dražan, S., and Brim, L. (2014). Robustness analysis of stochastic biochemical systems. *PLoS ONE* 9:e94553. doi: 10.1371/journal.pone.0094553

Chen, M., Cao, Y., and Watson, L. T. (2017). Parameter estimation of stochastic models based on limited data. *ACM SIG Bioinformatics* 7, 3:1–3:3. doi: 10.1145/3183624.3183627

Chevaliera, M. W., and Samadb, H. (2011). A data-integrated method for analyzing stochastic biochemical networks. *Chem. Phy.* 135:214110. doi: 10.1063/1.3664126

Chong, C. K., Mohamad, M. S., Deris, S., Shamsir, M. S., Chai, L. E., and Choon, Y. W. (2014). Parameter estimation by using an improved bee memory differential evolution algorithm (IBMDE) to simulate biochemical pathways. *Curr. Bioinform.* 9, 65–75. doi: 10.2174/15748936113080990007

Chong, C. K., Mohamad, M. S., Deris, S., Shamsir, M. S., Choon, Y. W., and Chai, L. E. (2012). Improved differential evolution algorithm for parameter estimation to improve the production of biochemical pathway. *Int. J. AI Interac. Multimed.* 1, 22–29. doi: 10.9781/ijimai.2012.153

Chou, I.-C., Martens, H., and Voit, E. O. (2006). Parameter estimation in biochemical systems models with alternating regression. *Theor. Biol. Med. Model.* 3:25. doi: 10.1186/1742-4682-3-25

Chou, I.-C., and Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.* 219, 57–83. doi: 10.1016/j.mbs.2009.03.002

Cseke, B., Schnoerr, D., Opper, M., and Sanguinetti, G. (2016). Expectation propagation for continuous time stochastic processes. *J. Phys.* 49:494002. doi: 10.1088/1751-8113/49/49/494002

Dai, Z., and Lai, L. (2010). Differential simulated annealing: a robust and efficient global optimization algorithm for parameter estimation of biological networks. *Mol. BioSyst.* 10, 1385–1392. doi: 10.1039/c4mb00100a

Daigle, B. J., Roh, M. K., Petzold, L. R., and Niemi, J. (2012). Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinf.* 13:68. doi: 10.1186/1471-2105-13-68

Dargatz, C. (2010). *Bayesian inference for diffusion processes with applications in life sciences* (PhD thesis), Ludwig Maximilians Universität München.

Dattner, I. (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Stat.* 9, 1939–1973. doi: 10.1214/15-EJS1053

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Rev. Microbiol.* 8, 717–729. doi: 10.1038/nrmicro2419

Deng, Z., and Tian, T. (2014). A continuous optimization approach for inferring parameters in mathematical models of regulatory networks. *BMC Bioinf.* 15:256. doi: 10.1186/1471-2105-15-256

Dey, A., Chakrabarti, K., Gola, K. K., and Sen, S. (2018). A Kalman filter approach for biomolecular systems with noise covariance updating. arXiv:1712.02150v2 [q-bio.QM].

Dochain, D. (2003). State observers for processes with uncertain kinetics. *Int. J. Control* 76, 1483–1492. doi: 10.1080/00207170310001604936

Doucet, A., Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. New York, NY: Springer-Verlag.

Drovandi, C. C., Pettitt, A. N., and Mccutchan, R. A. (2016). Exact and approximate Bayesian inference for low integer-valued time series models with intractable likelihoods. *Bayesian Anal.* 11, 325–352. doi: 10.1214/15-BA950

Eghtesadi, Z., and Mcauley, K. B. (2014). Mean square error based method for parameter ranking and selection to obtain accurate predictions at specified operating conditions. *Ind. Eng. Chem. Res.* 53, 6033–6046. doi: 10.1021/ie5002444

Eisenberg, M. C., and Hayashi, M. A. (2014). Determining identifiable parameter combinations using subset profiling. *Math. Biosci.* 256, 116–126. doi: 10.1016/j.mbs.2014.08.008

Emmert-Streib, F., Glazko, G. V., Altay, G., and de Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* 3:8. doi: 10.3389/fgene.2012.00008

Engl, H. W., Flamm, C., Kügler, P., Lu, J., Müller, S., and Schuster, P. (2009). Inverse problems in systems biology. *Inverse Probl.* 25, 1–51. doi: 10.1088/0266-5611/25/12/123014

Erguler, K., and Stumpf, M. P. (2011). Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol. Biosyst.* 7, 1593–1602. doi: 10.1039/c0mb00107d

Fages, F., Gay, S., and Soliman, S. (2015). Inferring reaction systems from ordinary differential equations. *Theor. Comput. Sci.* 599, 64–78. doi: 10.1016/j.tcs.2014.07.032

Famili, I., Mahadevan, R., and Palsson, B. O. (2005). K-cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophysics* 88, 1616–1625. doi: 10.1529/biophysj.104.050385

Farza, I. B. M., Menard, T., Abdennour, R. B., and M'Saad, M. (2016). On-line estimation of the reaction rates from sampled measurements in bioreactors. *IFAC Symp.* 49, 1205–1210. doi: 10.1016/j.ifacol.2016.07.375

Fearnhead, P., Giagos, V., and Sherlock, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics.* 70, 457–466. doi: 10.1111/biom.12152

Fearnhead, P., and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. B* 74, 419–474. doi: 10.1111/j.1467-9868.2011.01010.x

Fey, D., and Bullinger, E. (2010). "Limiting the parameter search space for dynamic models with rational kinetics using semi-definite programming," in *CAB Symp* (Leuven), 1–6.

Fey, D., Findeisen, R., and Bullinger, E. (2008). Parameter estimation in kinetic reaction models using nonlinear observers is facilitated by model extensions. *IFAC* 41, 313–318. doi: 10.3182/20080706-5-KR-1001.3521

Flassig, R. J. (2014). *Statistical model identification: dynamical processes and large-scale networks in systems biology* (PhD thesis), Otto-von-Guericke-Universität Magdeburg.

Folia, M. M., and Rattray, M. (2018). Trajectory inference and parameter estimation in stochastic models with temporally aggregated data. *Stat. Comput.* 28, 1053–1072. doi: 10.1007/s11222-017-9779-x

Fröhlich, F., Hross, S., Theis, F. J., and Hasenauer, J. (2014). "Radial basis function approximations of Bayesian parameter posterior densities for uncertainty analysis," in *CMSB* (Cham: Springer), 73–85.

Fröhlich, F., Kaltenbacher, B., Theis, F. J., and Hasenauer, J. (2017). Scalable parameter estimation for genomescale biochemical reaction networks. *PLOS Comput. Biol.* 13:e1005331. doi: 10.1371/journal.pcbi.1005331

Fröhlich, F., Thomas, P., Kazeroonian, A., Theis, F. J., Grima, R., and Hasenauer, J. (2016). Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.* 7:e1005030. doi: 10.1371/journal.pcbi.1005030

Gábor, A., and Banga, J. R. (2014). "Improved parameter estimation in kinetic models: selection and tuning of regularization methods," in *CMSB* (Cham: Springer), 45–60.

Gábor, A., Villaverde, A. F., and Banga, J. R. (2017). Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Syst. Biol.* 11:54. doi: 10.1186/s12918-017-0428-y

Galagali, N. (2016). *Bayesian Inference of Chemical Reaction Networks* (PhD thesis), Massachusetts Institute of Technology.

Geffen, D., Findeisen, R., Schliemann, M., Allgöwer, F., and Guay, M. (2008). "Observability based parameter identifiability for biochemical reaction networks," in *ACC* (Seattle, WA), 2130–2135.

Gennemark, P., and Wedelin, D. (2007). Efficient algorithms for ordinary differential equation model identification of biological systems. *IET Syst. Biol.* 1, 120–129. doi: 10.1049/iet-syb:20050098

Ghusinga, K. R., Vargas-Garcia, C. A., Lamperski, A., and Singh, A. (2017). Exact lower and upper bounds on stationary moments in stochastic biochemical systems. *Phys. Biol.* 14:eaa75c6. doi: 10.1088/1478-3975/aa75c6

Gillespie, C. S., and Golightly, A. (2012). "Bayesian inference for the chemical master equation using approximate models," in *WCSB* (Tampere), 23–26.

Golightly, A., Bradley, E., Lowe, T., and Gillespie, C. S. (2019). Correlated pseudo-marginal schemes for time-discretised stochastic kinetic models. *Comput. Stat. Data Anal.* 136, 92–107. doi: 10.1016/j.csda.2019.01.006

Golightly, A., Henderson, D. A., and Sherlock, C. (2012). Efficient particle MCMC for exact inference in stochastic biochemical network models through approximation of expensive likelihoods. *Stat. Comput.* 25, 1039–1055.

Golightly, A., Henderson, D. A., and Sherlock, C. (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Stat. Comput.* 25, 1039–1055. doi: 10.1007/s11222-014-9469-x

Golightly, A., and Kypraios, T. (2017). Efficient SMC2 schemes for stochastic kinetic models. *Stat. Comput.* 28, 1215–1230. doi: 10.1007/s11222-017-9789-8

Golightly, A., and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* 61, 781–788. doi: 10.1111/j.1541-0420.2005.00345.x

Golightly, A., and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov Chain Monte Carlo. *Interf. Focus* 1, 807–820. doi: 10.1098/rsfs.2011.0047

Gordon, N., Salmond, D., and Smith, A. (1993). A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Pro.-F* 140, 107–113. doi: 10.1049/ip-f-2.1993.0015

Goutsias, J., and Jenkinson, G. (2013). Markovian dynamics on complex reaction networks. *Phys. Rep.* 529, 199–264. doi: 10.1016/j.physrep.2013.03.004

Grant, M. J., and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *J. Health Info. Lib.* 26, 91–108. doi: 10.1111/j.1471-1842.2009.00848.x

Gratie, D.-E., Iancu, B., and Petre, I. (2013). "ODE analysis of biological systems," in *Formal Methods for Dynamical Systems* (Berlin; Heidelberg: Springer), 29–62.

Guillén-Gosálbez, G., Miró, A., Alves, R., Sorribas, A., and Jiménez, L. (2013). Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization. *BMC Syst. Biol.* 7:113. doi: 10.1186/1752-0509-7-113

Gupta, A. (2013). *Parameter estimation in deterministic and stochastic models of biological systems* (PhD thesis), University of Wisconsin-Madison.

Gupta, A., and Rawlings, J. B. (2014). Comparison of parameter estimation methods in stochastic chemical kinetic models: examples in systems biology. *AIChE* 60, 1253–1268. doi: 10.1002/aic.14409

Hagen, D. R., White, J. K., and Tidor, B. (2013). Convergence in parameters and predictions using computational experimental design. *Interf. Focus* 3:20130008. doi: 10.1098/rsfs.2013.0008

Hasenauer, J. (2013). *Modeling and Parameter Estimation for Heterogeneous Cell Populations* (PhD thesis), University of Stuttgart.

Hasenauer, J., Waldherr, S., Wagner, K., and Allgöwer, F. (2010). Parameter identication, experimental design and model falsication for biological network models using semidenite programming. *IET Syst. Biol.* 4, 119–130. doi: 10.1049/iet-syb.2009.0030

He, J., Sosonkina, M., Shaffer, C. A., Tyson, J. J., Watson, L. T., and Zwolak, J. W. (2004). "A hierarchical parallel scheme for global parameter estimation in systems biology," in *Proceedings - International Parallel and Distributed Processing Symposium* (Santa Fe, NM), 42–50.

Hori Y, Khammash MH, Hara S. (2013). "Efficient parameter identification for stochastic biochemical networks using a reduced-order realization," in *ECC* (Zurich), 4154–4159.

Hussain, F. (2016). *Techniques for automated parameter estimation in computational models of probabilistic systems* (PhD thesis), University of Central Florida.

Hussain, F., Langmead, C. J., Mi, Q., Dutta-Moscato, J., Vodovotz, Y., and Jha, S. K. (2015). Automated parameter estimation for biological models using Bayesian statistical model checking. *BMC Bioinformatics* 16:S8. doi: 10.1186/1471-2105-16-S17-S8

Iwata, M., Sriyudthsak, K., Hirai, M. Y., and Shiraishi, F. (2014). Estimation of kinetic parameters in an S-system equation model for a metabolic reaction system using the Newton-Raphson method. *Math. Biosci.* 248, 11–21. doi: 10.1016/j.mbs.2013.11.002

Jagiella, N., Rickert, D., Theis, F. J., and Hasenauer, J. (2017). Parallelization and high-performance computing enables automated statistical inference of multiscale models. *Cell Syst.* 4, 194–206. doi: 10.1016/j.cels.2016.12.002

Jang, H., Kim, K.-K. K., Braatz, R. D., Gopaluni, R. B., and Lee, J. H. (2016). Regularized maximum likelihood estimation of sparse stochastic monomolecular biochemical reaction networks. *Comput. Chem. Eng.* 90, 111–120. doi: 10.1016/j.compchemeng.2016.03.018

Jaqaman, K., and Danuser, G. (2006). Linking data to models: data regression. *Nat. Rev. Mol. Cell Biol.* 7, 813–819. doi: 10.1038/nrm2030

Ji, Z., and Brown, M. (2009). Joint state and parameter estimation for biochemical dynamic pathways with iterative extended Kalman filter: comparison with dual state and parameter estimation. *Open Aut. Contr. Syst.* 2, 69–77. doi: 10.2174/1874444300902010069

Jia, G., Stephanopoulos, G. N., and Gunawan, R. (2011). Parameter estimation of kinetic models from metabolic profiles: two-phase dynamic decoupling method. *Syst. Biol.* 27, 1964–1970. doi: 10.1093/bioinformatics/btr293

Joshia, M., Seidel-Morgenstern, A., and Kremling, A. (2006). Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metab. Eng.* 8, 447–455. doi: 10.1016/j.ymben.2006.04.003

Karimi, H., and Mcauley, K. B. (2013). An approximate expectation maximization algorithm for estimating parameters, noise variances, and stochastic disturbance intensities in nonlinear dynamic models. *Ind. Eng. Chem. Res.* 52, 18303–18323. doi: 10.1021/ie4023989

Karimi, H., and Mcauley, K. B. (2014a). An approximate expectation maximisation algorithm for estimating parameters in nonlinear dynamic models with process disturbances. *Can. J. Chem. Eng.* 92, 835–850. doi: 10.1002/cjce.21932

Karimi, H., and Mcauley, K. B. (2014b). A maximum-likelihood method for estimating parameters, stochastic disturbance intensities and measurement noise variances in nonlinear dynamic models with process disturbances. *Comp. Chem. Eng.* 67, 178–198. doi: 10.1016/j.compchemeng.2014.04.007

Karnaukhov, A. V., Karnaukhova, E. V., and Williamson, J. R. (2007). Numerical matrices method for nonlinear system identification and description of dynamics of biochemical reaction networks. *Biophysics* 92, 3459–3473. doi: 10.1529/biophysj.106.093344

Kimura, A., Celani, A., Nagao, H., Stasevich, T., and Nakamura, K. (2015). Estimating cellular parameters through optimization procedures: elementary principles and applications. *Front. Physiol.* 6:60. doi: 10.3389/fphys.2015.00060

Kleinstein, S. H., Bottino, D., Georgieva, A., and Sarangapani, R. (2006). "Nonuniform sampling for global optimization of kinetic rate constants in biological pathways," in *Winter Simulation Conference* (Monterey, CA), 1611–1616.

Ko, C. L., Voit, E. O., and Wang, F.-S. (2009). Estimating parameters for generalized mass action models with connectivity information. *BMC Bioinformatics* 10:140. doi: 10.1186/1471-2105-10-140

Koblents, E., Mariño, I. P., and Míguez, J. (2019). Bayesian computation methods for inference in stochastic kinetic models. *Complexity* 2019:7160934. doi: 10.1155/2019/7160934

Koblents, E., and Míguez, J. (2011). "A population Monte Carlo method for Bayesian inference and its application to stochastic kinetic models," in *EUSIPCO* (Barcelona), 679–683.

Koblents, E., and Míguez, J. (2014). A comparison of nonlinear population Monte Carlo and particle Markov Chain Monte Carlo algorithms for Bayesian inference in stochastic kinetic models. arXiv:1404.5218v1 [stat.ME].

Koeppl, H., Setti, G., Pelet, S., Mangia, M., Petrov, T., and Peter, M. (2010). "Probability metrics to calibrate stochastic chemical kinetics," in *ISCS* (Paris), 541–544.

Koeppl, H., Zechner, C., Ganguly, A., Pelet, S., and Peter, M. (2012). Accounting for extrinsic variability in the estimation of stochastic rate constants. *Int. J. Robust. Nonlinear Contr.* 22, 1103–1119. doi: 10.1002/rnc.2804

Komorowski, M., Costa, M. J., Rand, D. A., and Stumpf, M. P. (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *PNAS* 108, 8645–8650. doi: 10.1073/pnas.1015814108

Kravaris, C., Hahn, J., and Chu, Y. (2013). Advances and selected recent developments in state and parameter estimation. *Comput. Chem. Eng.* 51, 111–123. doi: 10.1016/j.compchemeng.2012.06.001

Kuepfer, L., Sauer, U., and Parrilo, P. A. (2007). Methodology article efficient classification of complete parameter regions based on semidefinite programming. *BMC Bioinformatics* 8:12. doi: 10.1186/1471-2105-8-12

Kügler, P. (2012). Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. *PLoS ONE* 7:e43001. doi: 10.1371/journal.pone.0043001

Kulikov, G., and Kulikova, M. (2017). Accurate state estimation of stiff continuous-time stochastic models in chemical and other engineering. *Math. Comput. Simul.* 142, 62–81. doi: 10.1016/j.matcom.2017.04.006

Kulikov, G. Y., and Kulikova, M. V. (2015a). High-order accurate continuous-discrete extended Kalman lter for chemical engineering. *Eur. J. Control.* 21, 14–26. doi: 10.1016/j.ejcon.2014.11.003

Kulikov, G. Y., and Kulikova, M. V. (2015b). "State estimation in chemical systems with infrequent measurements," in *ECC* (Linz), 2688–2693.

Kuntz, J., Thomas, P., Stan, G.-B., and Barahona, M. (2017). Rigorous bounds on the stationary distributions of the chemical master equation via mathematical programming. arXiv:1702.05468v1 [math.PR].

Kurt, Z., Aydin, N., and Altay, G. (2016). Comprehensive review of association estimators for the inference of gene networks. *Turkish J. EECS* 24, 695–718. doi: 10.3906/elk-1312-90

Kutalik, Z., Tucker, W., and Moulton, V. (2007). S-system parameter estimation for noisy metabolic profiles using Newton-flow analysis. *IET Syst. Biol.* 1, 174–180. doi: 10.1049/iet-syb:20060064

Kuwahara, H., Fan, M., Wang, S., and Gao, X. (2013). A framework for scalable parameter estimation of gene circuit models using structural information. *ISMB/ECCB* 29, i98–i107. doi: 10.1093/bioinformatics/btt232

Kyriakopoulos, C., and Wolf, V. (2015). "Optimal observation time points in stochastic chemical kinetics," in *Hybrid Systems Biology* (Cham: Springer), 83–96.

Lakatos, E. (2017). *Stochastic analysis and control methods for molecular cell biology* (PhD thesis), Imperial College London.

Lakatos, E., Ale, A., Kirk, P. D., and Stumpf, M. P. (2015). Multivariate moment closure techniques for stochastic kinetic models. *Chem. Phys.* 143, 1–13. doi: 10.1063/1.4929837

Lang, M., and Stelling, J. (2016). Modular parameter identification of biomolecular networks. *SIAM J. Sci. Comput.* 38, B988–B1008. doi: 10.1137/15M103306X

Lecca, P., Palmisano, A., Priami, C., and Sanguinetti, G. (2009). "A new probabilistic generative model of parameter inference in biochemical networks," in *SAC* (Honolulu, HI), 758–765.

Li, P., and Vu, Q. D. (2013). Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC Syst. Biol.* 7:91. doi: 10.1186/1752-0509-7-91

Li, P., and Vu, Q. D. (2015). A simple method for identifying parameter correlations in partially observed linear dynamic models. *BMC Syst. Biol.* 9:3. doi: 10.1186/s12918-015-0234-3

Liao, S. (2017). *High-dimensional problems in stochastic modelling of biological processes* (PhD thesis), University of Oxford.

Liao, S., Vejchodský, T., and Erban, R. (2015). Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *J. R. Soc. Interface* 12:20150233. doi: 10.1098/rsif.2015.0233

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* 9, 439–456. doi: 10.1038/nprot.2014.025

Lillacci, G., and Khammash, M. (2010). "Parameter identification of biological networks using extended Kalman filtering and χ^2 criteria," in *CDC* (Atlanta, GA), 3367–3372.

Lillacci, G., and Khammash, M. (2012). A distribution-matching method for parameter estimation and model selection in computational biology. *Int. J. Robust Nonlinear Contr.* 22, 1065–1081. doi: 10.1002/rnc.2794

Linder, D. F. (2013). *Penalized least squares and the algebraic statistical model for biochemical reaction networks* (PhD thesis), Georgia Regents University, St, Augusta, GA, United States.

Lindera, D. F., and Rempala, G. A. (2015). Bootstrapping least-squares estimates in biochemical reaction networks. *Biol. Dyn.* 9, 125–146. doi: 10.1080/17513758.2015.1033022

Liu, F., Farza, M., and Saad, M. M. (2006). Unknown input observers design for a class of nonlinear systems–application to biochemical processes. *ROCOND* 39, 131–136. doi: 10.3182/20060705-3-FR-2907.00024

Liu, L.-Z., Wu, F.-X., and Zhang, W. J. (2012). Inference of biological S-system using the separable estimation method and the genetic algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* 9, 955–965. doi: 10.1109/TCBB.2011.126

Liu, P.-K., and Wang, F.-S. (2008a). Inference of biochemical network models in S-system using multiobjective optimization approach. *Syst. Biol.* 24, 1085–1092. doi: 10.1093/bioinformatics/btn075

Liu, P. K., and Wang, F.-S. (2008b). Inverse problems of biological systems using multi-objective optimization. *J. Chinese Inst. Chem. Eng.* 39, 399–406. doi: 10.1016/j.jcice.2008.05.001

Liu, P. K., and Wang, F. S. (2009). Hybrid differential evolution with geometric mean mutation in parameter estimation of bioreaction systems with large parameter search space. *Comput. Chem. Eng.* 33, 1851–1860. doi: 10.1016/j.compchemeng.2009.05.008

Liu, X. (2014). *Probabilistic inference in models of systems biology* (PhD thesis), University of Southampton, Southampton, United Kingdom.

Liu, Y., and Gunawan, R. (2014). Parameter estimation of dynamic biological network models using integrated fluxes. *BMC Syst. Biol.* 8:127. doi: 10.1186/s12918-014-0127-x

Loos, C., Fiedler, A., and Hasenauer, J. (2016). "Parameter estimation for reaction rate equation constrained mixture models," in *CMSB* (Cham: Springer), 186–200.

Loskot, P. (2018). Automation is coming to research. *IEEE Sig. Proc. Mag.* 35, 138–140. doi: 10.1109/MSP.2018.2811006

Loskot, P., Atitey, K., and Mihaylova, L. (2019). Comprehensive review of models and methods for inferences in bio-chemical reaction networks. arXiv:1902.05828 [q-bio.QM].

Lötstedt, P. (2018). The linear noise approximation for spatially dependent biochemical networks. *Bull. Math. Biol.* 81, 1–29. doi: 10.1007/s11538-018-0428-0

Lück, A., and Wolf, V. (2016). Generalized method of moments for estimating parameters of stochastic reaction networks. *BMC Syst. Biol.* 10:8. doi: 10.1186/s12918-016-0342-8

Mancini, T., Tronci, E., Salvo, I., Mari, F., Massini, A., and Melatti, I. (2015). "Computing biological model parameters by parallel statistical model checking," in *IWBBIO* (Granada), 542–554.

Mannakee, B. K., Ragsdale, A. P., Transtrum, M. K., and Gutenkunst, R. N. (2016). "Sloppiness and the geometry of parameter space," in *Uncertainty in Biology. Studies in Mechanobiology, Tissue Engineering and Biomaterials* (Cham: Springer), 271–299.

Mansouri, M., Avci, O., Nounou, H., and Nounou, M. (2015). "Parameter identification for nonlinear biological phenomena modeled by S-systems," in *SSD* (Mahdia), 1–6.

Mansouri, M. M., Nounou, H. N., Nounou, M. N., and Datta, A. A. (2014). Modeling of nonlinear biological phenomena modeled by S-systems. *Math. Biosci.* 249, 75–91. doi: 10.1016/j.mbs.2014.01.011

Matsubara, Y., Kikuchi, S., Sugimoto, M., and Tomita, M. (2006). Parameter estimation for stiff equations of biosystems using radial basis function networks. *BMC Bioinf.* 7:230. doi: 10.1186/1471-2105-7-230

Mazur, J. (2012). *Bayesian inference of gene regulatory networks: from parameter estimation to experimental design* (PhD thesis), Ruprecht-Karls Ruprecht-Karls-Universität, Heidelberg, Germany.

Mazur, J., and Kaderali, L. (2013). "The importance and challenges of bayesian parameter learning in systems biology," in *Model Based Parameter Estimation* (Berlin: Springer-Verlag), 145–156.

McGoff, K., Mukherjee, S., and Pillai, N. (2015). Statistical inference for dynamical systems: a review. *Stat. Surv.* 9, 209–252. doi: 10.1214/15-SS111

Mendes, P., and Kell, D. (1998). Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869–883.

Meskin, N., Nounou, H., Nounou, M., Datta, A., and Dougherty, E. R. (2011). "Parameter estimation of biological phenomena modeled by S-systems: an extended Kalman filter approach," in *CDC-ECC* (Orlando, FL), 4424–4429.

Michailidis, G., and d'Alché-Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Math. Biosci.* 246, 326–334. doi: 10.1016/j.mbs.2013.10.003

Michalik, C., Chachuat, B., and Marquardt, W. (2009). Incremental global parameter estimation in dynamical systems. *Ind. Eng. Chem. Res.* 48, 5489–5497. doi: 10.1021/ie8015472

Mihaylova, L., Carmi, A. Y., Septier, F., Gning, A., Pang, S. K., and Godsill, S. (2014). Overview of Bayesian sequential Monte Carlo methods for group and extended object tracking. *Dig. Signal Proces.* 25, 1–16. doi: 10.1016/j.dsp.2013.11.006

Mihaylova, L., Hegyi, A., Gning, A., and Boel, R. (2012). Parallelized particle and Gaussian sum particle filters for large scale freeway traffic systems. *IEEE Trans. Intel. Transport. Syst.* 13, 36–48. doi: 10.1109/TITS.2011.2178833

Mikeev, L., and Wolf, V. (2012). "Parameter estimation for stochastic hybrid models of biochemical reaction networks," in *HSCC* (Beijing), 155–166.

Mikelson, J., and Khammash, M. (2016). A parallelizable sampling method for parameter inference of large biochemical reaction models. arXiv:1606.08281v1 [q-bio.QM].

Milios, D., Sanguinetti, G., and Schnoerr, D. (2018). "Probabilistic model checking for continuous-time Markov chains via sequential Bayesian inference," in *QEST*, Vol. 11024 (Cham: Springer), 289–305.

Milner, P., Gillespie, C. S., and Wilkinson, D. J. (2013). Moment closure based parameter inference of stochastic kinetic models. *Stat. Comput.* 23, 287–295. doi: 10.1007/s11222-011-9310-8

Mizera, A., Pang, J., and Yuan, Q. (2014). "Model-checking based approaches to parameter estimation of gene regulatory networks," in *International Conference on Engineering of Complex Computer Systems* (Tianjin), 206–209.

Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13, 2467–2474. doi: 10.1101/gr.1262503

Moreno, J., and Denis, D. (2005). "Global observability and detectability analysis of uncertain reaction systems," in *IFAC World Congress* 1062–1070.

Moritz, L. (2014). *Modular identification and analysis of biomolecular networks* (PhD thesis), ETH Zurich.

Mozgunov, P., Beccuti, M., Horvath, A., Jaki, T., Sirovich, R., and Bibbona, E. (2018). A review of the deterministic and diffusion approximations for stochastic chemical reaction networks. *Reac. Kinet. Mech. Cat.* 123, 289–312. doi: 10.1007/s11144-018-1351-y

Mu, L. (2010). *Parameter estimation methods for biological systems* (PhD thesis), University of Saskatchewan.

Müller, C. L., Ramaswamy, R., and Sbalzarini, I. F. (2012). Global parameter identification of stochastic reaction networks from single trajectories. *Adv. Exp. Med. Biol.* 736, 477–498. doi: 10.1007/978-1-4419-7210-128

Murakami, Y. (2014). Bayesian parameter inference and model selection by population annealing in systems biology. *PLoS ONE* 9:e104057. doi: 10.1371/journal.pone.0104057

Murtuza Baker, S. M., Poskar, C. H., Schreiber, F., and Junker, B. H. (2013). An improved constraint filtering technique for inferring hidden states and parameters of a biological model. *Bioinformatics* 29, 1052–1059. doi: 10.1093/bioinformatics/btt097

Nemeth, C., Fearnhead, P., and Mihaylova, L. (2014). Sequential Monte Carlo methods for state and parameter estimation in abruptly changing environments. *IEEE Trans. Signal Proces.* 62, 1245–1255. doi: 10.1109/TSP.2013.2296278

Nienałtowski, K., Włodarczyk, M., Lipniacki, T., and Komorowski, M. (2015). Clustering reveals limits of parameter identifiability in multi-parameter models of biochemical dynamics. *BMC Syst. Biol.* 9:8. doi: 10.1186/s12918-015-0205-8

Nim, T. H., Luo, L., Clément, M.-V., White, J. K., and Kellogg, L. (2013). Systematic parameter estimation in data-rich environments for cell signalling dynamics. *Syst. Biol.* 29, 1044–1051. doi: 10.1093/bioinformatics/btt083

Nobile, M. S., Besozzi, D., Cazzaniga, P., Mauri, G., and Pescini, D. (2012). "A GPU-based multi-swarm PSO method for parameter estimation in stochastic biological systems exploiting discrete-time target series," in *EvoBIO 2012*. LNCS 7246 (Berlin; Heidelberg: Springer), 74–85.

Nobile, M. S., Besozzi, D., Cazzaniga, P., Pescini, D., and Mauri, G. (2013). "Reverse engineering of kinetic reaction networks by means of cartesian genetic programming and particle swarm optimization," in *IEEE Congress on Evolutionary Computation* (Cancun), 1594–1601.

Nobile, M. S., Cazzaniga, P., Besozzi, D., Colombo, R., Mauri, G., and Pasi, G. (2018a). Fuzzy self-tuning PSO: a settings-free algorithm for global optimization. *Swarm Evol. Comput.* 39, 70–85. doi: 10.1016/j.swevo.2017.09.001

Nobile, M. S., Pasi, G., Cazzaniga, P., Besozzi, D., Colombo, R., and Mauri, G. (2015). "Proactive particles in swarm optimization: a self-tuning algorithm based on fuzzy logic," in *FUZZ-IEEE* (Istanbul), 1–8.

Nobile, M. S., Tangherloni, A., Besozzi, D., and Cazzaniga, P. (2016). "GPU-powered and settings-free parameter estimation of biochemical systems," in *CEC* (Vancouver, BC), 32–39.

Nobile, M. S., Tangherloni, A., Rundo, L., Spolaor, S., Besozzi, D., Mauri, G., and Cazzaniga, P. (2018b). "Computational intelligence for parameter estimation of biochemical systems," in *IEEE CEC* (Rio de Janeiro), 1–8.

Onwuegbuzie, A. J., and Frels, R. (2016). *7 Steps to a Comprehensive Literature Review: A Multimodal and Cultural Approach.* London: Sage Public.

Pahle, J., Challenger, J. D., Mendes, P., and McKane, A. J. (2012). Biochemical fluctuations, optimisation and the linear noise approximation. *BMC Syst. Biol.* 6:86. doi: 10.1186/1752-0509-6-86

Palmisano, A. (2010). *Modelling and inference strategies for biological systems* (PhD thesis), Università degli Studi di Trento.

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pantazis, Y., Katsoulakis, M. A., and Vlachos, D. G. (2013). Parametric sensitivity analysis for biochemical reaction networks based on pathwise information theory. *BMC Bioinf.* 14, 1–19. doi: 10.1186/1471-2105-14-311

Paul, D. (2014). *Efficient parameter inference for stochastic chemical kinetics* (PhD thesis), Uppsala University.

Penas, D. R., González, P., Egea, J. A., Doallo, R., and Banga, J. R. (2017). Parameter estimation in large-scale systems biology models: a parallel and self-adaptive cooperative strategy. *BMC Bioinformatics* 18:4. doi: 10.1186/s12859-016-1452-4

Plesa, T., Vejchodský, T., and Erban, R. (2017). "Test models for statistical inference: two-dimensional reaction systems displaying limit cycle bifurcations and bistability," in *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology* (Cham: Springer), 3–27.

Pullen, N., and Morris, R. J. (2014). Bayesian model comparison and parameter inference in systems biology using nested sampling. *PLoS ONE* 9:e88419. doi: 10.1371/journal.pone.0088419

Radulescu, Gorban, A. N., Zinovyev, A., and Noel, V. (2012). Reduction of dynamical biochemical reactions networks in computational biology. *Front. Genet.* 3:131. doi: 10.3389/fgene.2012.00131

Rakhshania, H., Dehghanianb, E., and Rahatia, A. (2016). Hierarchy cuckoo search algorithm for parameter estimation in biological systems. *Chemomet. Intel. Lab. Syst.* 159, 97–107. doi: 10.1016/j.chemolab.2016.10.011

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. B* 69, 741–796. doi: 10.1111/j.1467-9868.2007.00610.x

Rapaport, A., and Dochain, D. (2005). Interval observers for biochemical processes with uncertain kinetics and inputs. *Math. Biosci.* 193, 235–253. doi: 10.1016/j.mbs.2004.07.004

Reinker, S., Altman, R., and Timmer, J. (2006). Parameter estimation in stochastic biochemical reactions. *IEE Proc. Syst. Biol.* 153, 168–178. doi: 10.1049/ip-syb:20050105

Reis, M., Kromer, J. A., and Klipp, E. (2018). General solution of the chemical master equation and modality of marginal distributions for hierarchic first-order reaction networks. *J. Math. Biol.* 77, 377–419. doi: 10.1007/s00285-018-1205-2

Remlia, M. A., Derisb, S., Mohamada, M. S., Omatuc, S., and Corchadod, J. M. (2017). An enhanced scatter search with combined opposition-based learning for parameter estimation in large-scale kinetic models of biochemical systems. *Eng. Appl. AI* 62, 164–180. doi: 10.1016/j.engappai.2017.04.004

Rempala, G. A. (2012). Least squares estimation in stochastic biochemical networks. *Bull. Math. Biol.* 74, 1938–1955. doi: 10.1007/s11538-012-9744-y

Revell, J., and Zuliani, P. (2018). "Stochastic rate parameter inference using the cross-entropy method," in *CMSB* (Cham: Springer), 146–164.

Rodriguez-Fernandez, M., Egea, J. A., and Banga, J. R. (2006a). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* 7:483. doi: 10.1186/1471-2105-7-483

Rodriguez-Fernandez, M., Mendes, P., and Banga, J. R. (2006b). A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSyst.* 83, 248–265. doi: 10.1016/j.biosystems.2005.06.016

Rodriguez-Fernandez, M., Rehberg, M., Kremling, A., and Banga, J. R. (2013). Simultaneous model discrimination and parameter estimation in dynamic models of cellular systems. *BMC Syst. Biol.* 7:76. doi: 10.1186/1752-0509-7-76

Rosati, E., Madec, M., Kammerer, J.-B., Hérard, L., Lallement, C., and Haiech, J. (2018). Efficient modeling and simulation of space-dependent biological systems. *Comput. Biol.* 25, 917–933. doi: 10.1089/cmb.2018.0012

Ruess, J. (2014). *Moment-based methods for the analysis and identification of stochastic models of biochemical reaction networks* (PhD thesis), ETH.

Ruess, J., and Lygeros, J. (2015). Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM Trans. Model. Comput. Simul.* 25:8. doi: 10.1145/2688906

Ruess, J., Milias-Argeitis, A., Summers, S., and Lygeros, J. (2011). Moment estimation for chemically reacting systems by extended Kalman filtering. *Chem. Phys.* 135:165102. doi: 10.1063/1.3654135

Rumschinski, P., Borchers, S., Bosio, S., Weismantel, R., and Findeisen, R. (2010). Set-base dynamical parameter estimation and model invalidation for biochemical reaction networks. *BMC Syst. Biol.* 4:69. doi: 10.1186/1752-0509-4-69

Ruttor, A., and Opper, M. (2009). Efficient statistical inference for stochastic reaction processes. *Phys. Rev. Lett.* 103:230601. doi: 10.1103/PhysRevLett.103.230601

Sadamoto, T., Ishizaki, T., and ichi Imura, J. (2017). Average state observers for large-scale network systems. *IEEE Trans. Contr. Net. Syst.* 4, 761–769. doi: 10.1109/TCNS.2016.2550866

Saltelli, A., Ratto, M., Tarantola, S., and Campolongo, F. (2005). Sensitivity analysis for chemical models. *Chem. Rev.* 105, 2811–2828. doi: 10.1021/cr040659d

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity Analysis in Practice*. Chichester: Wiley.

Schenkendorf, R. (2014). *Optimal experimental design for parameter identification and model selection* (PhD thesis), Otto-von-Guericke-Universität Magdeburg.

Schilling, C., Bogomolov, S., Henzinger, T. A., Podelski, A., and Ruess, J. (2016). Adaptive moment closure for parameter inference of biochemical reaction networks. *BioSyst.* 149, 15–25. doi: 10.1016/j.biosystems.2016.07.005

Schnoerr, D. (2016). *Approximation methods and inference for stochastic biochemical kinetics* (PhD thesis), University of Edinburgh.

Schnoerr, D., Sanguinetti, G., and Grima, R. (2017). Approximation and inference methods for stochastic biochemical kinetics–a tutorial review. *J. Phys. A Math. Theor.* 50, 1–60. doi: 10.1088/1751-8121/aa54d9

Septier, F., and Peters, G. W. (2016). Langevin and Hamiltonian based sequential MCMC for efficient Bayesian filtering in high-dimensional spaces. *IEEE J. Sel. Topics Signal Proces.* 10, 312–327. doi: 10.1109/JSTSP.2015.2497211

Shacham, M., and Brauner, N. (2014). Application of stepwise regression for dynamic parameter estimation. *Comput. Chem. Eng.* 69, 26–38. doi: 10.1016/j.compchemeng.2014.06.013

Sherlock, C., Golightly, A., and Gillespie, C. S. (2014). Bayesian inference for hybrid discrete-continuous stochastic kinetic models. *Inv. Probl.* 30:114005. doi: 10.1088/0266-5611/30/11/114005

Shiang, K. D. (2009). A perturbation-based estimate algorithm for parameters of coupled ordinary differential equations, applications from chemical reactions to metabolic dynamics. *Comput. Method Prog. Biomed.* 94, 118–142. doi: 10.1016/j.cmpb.2008.12.001

Singh, A. K., and Hahn, J. (2005). State estimation for high-dimensional chemical processes. *Comput. Chem. Eng.* 29, 2326–2334. doi: 10.1016/j.compchemeng.2005.05.009

Slezak, D. F., Suárez, C., Cecchi, G. A., Marshall, G., and Stolovitzky, G. (2010). When the optimal is not the best: parameter estimation in complex biological models. *PLoS ONE* 5:e13283. doi: 10.1371/journal.pone.0013283

Smadbeck, P. (2014). *Chemical master equations for non-linear stochastic reaction networks: closure schemes and implications for discovery in the biological sciences* (PhD thesis), University Of Minnesota.

Smith, S., and Grima, R. (2018). Spatial stochastic intracellular kinetics: a review of modelling approaches. *Bull. Math. Biol.* doi: 10.1007/s11538-018-0443-1. [Epub ahead of print].

Srinath, S., and Gunawan, R. (2010). Parameter identifiability of power-law biochemical system models. *J. Biotechnol.* 149, 132–140. doi: 10.1016/j.jbiotec.2010.02.019

Srinivas, M., and Rangaiah, G. P. (2007). Differential evolution with tabu list for global optimization and its application to phase equilibrium and parameter estimation problems. *Ind. Eng. Chem. Res.* 46, 3410–3421. doi: 10.1021/ie0612459

Srivastava, R. (2012). *Parameter estimation in stochastic chemical kinetic models* (PhD thesis), University of Wisconsin-Madison.

Srivastavaa, R., and Rawlingsb, J. B. (2014). Parameter estimation in stochastic chemical kinetic models using derivative free optimization and bootstrapping. *Comput. Chem. Eng.* 63, 152–158. doi: 10.1016/j.compchemeng.2014.01.006

Sun, J., Garibaldi, J. M., and Hodgman, C. (2012). Parameter estimation using metaheuristics in systems biology: a comprehensive review. *ACM Trans. Comput. Biol. Bioinf.* 9, 185–202. doi: 10.1109/TCBB.2011.63

Sun, J., Palade, V., Cai, Y., Fang, W., and Wu, X. (2014). Biochemical systems identification by a random drift particle swarm optimization approach. *BMC Bioinf.* 15:S1. doi: 10.1186/1471-2105-15-S6-S1

Sun, X., Jin, L., and Xiong, M. (2008). Extended Kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks. *PLoS ONE* 3:e3758. doi: 10.1371/journal.pone.0003758

Tanevski, J., Džeroski, S., and Kocarev, L. (2010). Approximate Bayesian parameter inference for dynamical systems in systems biology. *Sec. Math. Tech. Sci.* XXXI, 73–98. doi: 10.20903/csnmbs.masa.2010.31.1-2.24

Tangherloni, A., Nobile, M. S., and Cazzaniga, P. (2016). "GPU-powered bat algorithm for the parameter estimation of biochemical kinetic values," in *CIBCB* (Chiang Mai), 1–6.

Teijeiro, D., Pardo, X. C., Penas, D. R., González, P., Banga, J. R., and Doallo, R. (2017). A cloud-based enhanced differential evolution algorithm for parameter estimation problems in computational systems biology. *Cluster Comput.* 20, 1937–1950. doi: 10.1007/s10586-017-0860-1

Tenazinha, N., and Vinga, S. (2011). A survey on methods for modeling and analyzing integrated biological networks. *Trans. Comput. Biol. Bioinform.* 8, 943–958. doi: 10.1109/TCBB.2010.117

Thomas, P., Straube, A. V., and Grima, R. (2012). The slow-scale linear noise approximation: An accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC Syst. Biol.* 6:39. doi: 10.1186/1752-0509-6-39

Tian, L.-P., Liu, L., and Wu, F. X. (2010). "Iterative linear least squares method of parameter estimation for linear-fractional models of molecular biological systems," in *ICBBE* (Chengdu), 1–4.

Tian, T., Xu, S., Gao, J., and Burrage, K. (2007). Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Syst. Biol.* 23, 84–91. doi: 10.1093/bioinformatics/btl552

Toni, T., and Stumpf, M. P. (2010). Parameter inference and model selection in signaling pathway models. *Methods Mol. Biol.* 673, 283–295. doi: 10.1007/978-1-60761-842-318

Tranfield, D., Denyer, D., and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manage.* 14, 207–222. doi: 10.1111/1467-8551.00375

Transtrum, M. K., and Qiu, P. (2012). Optimal experiment selection for parameter estimation in biological differential equation models. *BMC Bioinformatics* 13:181. doi: 10.1186/1471-2105-13-181

Vanlier, J., Tiemann, C. A., Hilbers, P. A., and van Riel, N. A. (2013). Parameter uncertainty in biochemical models described by ordinary differential equations. *Math. Biosci.* 246, 305–314. doi: 10.1016/j.mbs.2013.03.006

Vargas, A., Moreno, J., and Wouwer, A. V. (2014). A weighted variable gain super-twisting observer for the estimation of kinetic rates in biological systems. *J. Process Control* 24, 957–965. doi: 10.1016/j.jprocont.2014.04.018

Veerman, F., Popovic, N., and Marr, C. (2018). Parameter inference with analytical propagators for stochastic models of autoregulated gene expression. bioRxiv.

Venayak, N., von Kamp, A., Klamt, S., and Mahadevan, R. (2018). Move identifies metabolic valves to switch between phenotypic states. *Nat. Commun.* 9:5332. doi: 10.1038/s41467-018-07719-4

Villaverde, A. F., and Barreiro, A. (2016). Identifiability of large nonlinear biochemical networks. *MATCH Commun. Math. Comput. Chem.* 76, 259–296.

Villaverde, A. F., Barreiro, A., and Papachristodoulou, A. (2016). Structural identifiability of dynamic systems biology models. *PLoS Comput. Biol.* 12:e1005153. doi: 10.1371/journal.pcbi.1005153

Villaverde, A. F., Egea, J. A., and Banga, J. R. (2012). A cooperative strategy for parameter estimation in large scale systems biology models. *BMC Syst. Biol.* 6:75. doi: 10.1186/1752-0509-6-75

Villaverde, A. F., Ross, J., Morán, F., and Banga, J. R. (2014). MIDER: network inference with mutual information distance and entropy reduction. *PLoS ONE* 9:e96732. doi: 10.1371/journal.pone.0096732

Voit, E. O. (2013). Biochemical systems theory: a review. *ISRN Biomath.* 2013, 1–53. doi: 10.1155/2013/897658

von Stosch, M., Oliveira, R., Peres, J., and de Azevedo, S. F. (2014). Hybrid semi-parametric modeling in process systems engineering: past, present and future. *Comput. Chem. Eng.* 60, 86–101. doi: 10.1016/j.compchemeng.2013.08.008

Vrettas, M. D., Cornford, D., and Opper, M. (2011). Estimating parameters in stochastic systems: a variational Bayesian approach. *Phys. D* 240, 1877–1900. doi: 10.1016/j.physd.2011.08.013

Wang, Y., Christley, S., Mjolsness, E., and Xie, X. (2010). Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Syst. Biol.* 4:99. doi: 10.1186/1752-0509-4-99

Weber, M. F., and Frey, E. (2017). Master equations and the theory of stochastic path integrals. *Rep. Prog. Phys.* 80:046601. doi: 10.1088/1361-6633/aa5ae2

Weiss, K., Khoshgoftaar, T. M., and Wang, D.-D. (2016). A survey of transfer learning. *J. Big Data* 3, 1–40. doi: 10.1186/s40537-016-0043-6

Whitaker, G. A., Golightly, A., Boys, R. J., and Sherlock, C. (2017). Bayesian inference for diffusion-driven mixed-effects models. *Bayesian Anal.* 12, 435–463. doi: 10.1214/16-BA1009

White, A., Tolman, M., Thames, H. D., Withers, H. R., Mason, K. A., and Transtrum, M. K. (2016). The limitations of model-based experimental design and parameter estimation in sloppy systems. *PLoS Comput. Biol.* 12:e1005227. doi: 10.1371/journal.pcbi.1005227

White, S., Kypraios, T., and Preston, S. (2015). Piecewise approximate Bayesian computation: Fast inference for discretely observed Markov models using a factorised posterior distribution. *Stat. Comput.* 25, 289–301. doi: 10.1007/s11222-013-9432-2

Wong, M. K., Krycer, J. R., Burchfield, J. G., James, D. E., and Kuncic, Z. (2015). A generalised enzyme kinetic model for predicting the behaviour of complex biochemical systems. *FEBS Open Bio.* 5, 226–239. doi: 10.1016/j.fob.2015.03.002

Woodcock, D., Komorowski, M., Finkenstadt, B., Harper, C., Davis, J., White, M., and Rand, D. (2011). *A Bayesian Hierarchical Diffusion Model for Estimating Kinetic Parameters and Cell-to-Cell Variability*. Warwick University, 1–18.

Xiong, J., and Zhou, T. (2013). "Parameter identification for nonlinear state-space models of a biological network via linearization and robust state estimation," in *Chinese Control Conference* (Xi'an), 8235–8240.

Yang, X., Guo, Y., and Guo, L. (2014). An iterative parameter estimation method for biological systems and its parallel implementation. *Concur. Comput. Pract. Exp.* 26, 1249–1267. doi: 10.1002/cpe.3071

Yenkie, K., Diwekar, U., and Linninger, A. (2016). Simulation-free estimation of reaction propensities in cellular reactions and gene signaling networks. *Comput. Chem. Eng.* 87, 154–163. doi: 10.1016/j.compchemeng.2016.01.010

Zamora-Sillero, E., Hafner, M., Ibig, A., Stelling, J., and Wagner, A. (2011). Efficient characterization of high-dimensional parameter spaces for systems biology. *BMC Syt. Biol.* 5:142. doi: 10.1186/1752-0509-5-142

Zechner, C. (2014). *Stochastic biochemical networks in random environments: probabilistic modeling and inference* (PhD thesis), ETH.

Zechner, C., Nandy, P., Unger, M., and Koeppl, H. (2012). "Optimal variational perturbations for the inference of stochastic reaction dynamics," in *CDC* (Maui, HI), 5336–5341.

Zechner, C., Pelet, S., Peter, M., and Koeppl, H. (2011). "Recursive Bayesian estimation of stochastic rate constants from heterogeneous cell populations," in *CDC-ECC* (Orlando, FL), 5837–5843.

Zeng, N., Wang, Z., Li, Y., Du, M., and Liu, X. (2012). A hybrid EKF and switching PSO algorithm for joint state and parameter estimation of lateral flow immunoassay models. *Trans. Comput. Biol. Bioinform.* 9, 321–329. doi: 10.1109/TCBB.2011.140

Zhan, C., Situ, W., Yeung, L. F., Tsang, P. W.-M., and Yang, G. (2014). A parameter estimation method for biological systems modeled by ODEs/DDEs models using spline approximation and differential evolution algorithm. *ACM Trans. Comput. Biol. Bioinform.* 11, 1066–1076. doi: 10.1109/TCBB.2014.2322360

Zhan, C., and Yeung, L. F. (2011). Parameter estimation in systems biology models using spline approximation. *BMC Syst. Biol.* 5:14. doi: 10.1186/1752-0509-5-14

Zimmer, C. (2016). Experimental design for stochastic models of nonlinear signaling pathways using an interval-wise linear noise approximation and state estimation. *PLoS ONE* 11:e159902. doi: 10.1371/journal.pone.0159902

Zimmer, C., and Sahle, S. (2012). Parameter estimation for stochastic models of biochemical reactions. *Comput. Sci. Syst. Biol.* 6, 11–21. doi: 10.4172/jcsb.1000095

Zimmer, C., and Sahle, S. (2015). Deterministic inference for stochastic systems using multiple shooting and a linear noise approximation for the transition probabilities. *IET Syst. Biol.* 9, 181–192. doi: 10.1049/iet-syb.2014.0020

Zimmer, C., Sahle, S., and Pahle, J. (2014). Exploiting intrinsic fluctuations to identify model parameters. *IET Syst. Biol.* 9, 64–73. doi: 10.1049/iet-syb.2014.0010

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.

Copyright © 2019 Loskot, Atitey and Mihaylova. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 15 February 2019
doi: 10.3389/fgene.2019.00087

Construction of a Suite of Computable Biological Network Models Focused on Mucociliary Clearance in the Respiratory Tract

Hasmik Yepiskoposyan, Marja Talikka, Stefano Vavassori, Florian Martin, Alain Sewer, Sylvain Gubian, Karsta Luettich, Manuel Claude Peitsch and Julia Hoeng*

PMI R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

Mucociliary clearance (MCC), considered as a collaboration of mucus secreted from goblet cells, the airway surface liquid layer, and the beating of cilia of ciliated cells, is the airways' defense system against airborne contaminants. Because the process is well described at the molecular level, we gathered the available information into a suite of comprehensive causal biological network (CBN) models. The suite consists of three independent models that represent (1) cilium assembly, (2) ciliary beating, and (3) goblet cell hyperplasia/metaplasia and that were built in the Biological Expression Language, which is both human-readable and computable. The network analysis of highly connected nodes and pathways demonstrated that the relevant biology was captured in the MCC models. We also show the scoring of transcriptomic data onto these network models and demonstrate that the models capture the perturbation in each dataset accurately. This work is a continuation of our approach to use computational biological network models and mathematical algorithms that allow for the interpretation of high-throughput molecular datasets in the context of known biology. The MCC network model suite can be a valuable tool in personalized medicine to further understand heterogeneity and individual drug responses in complex respiratory diseases.

Keywords: mucociliary clearance, network models, biological expression language, respiratory tract, network perturbation amplitude

OPEN ACCESS

Edited by:
Marco Antoniotti, Università degli Studi di Milano Bicocca, Italy

Reviewed by:
Richard R. Rodrigues, Oregon State University, United States
Adriano Velasque Werhli, Fundação Universidade Federal do Rio Grande, Brazil

****Correspondence:***
Marja Talikka
marja.talikka@pmi.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *05 October 2018*
Accepted: *29 January 2019*
Published: *15 February 2019*

Citation:
Yepiskoposyan H, Talikka M, Vavassori S, Martin F, Sewer A, Gubian S, Luettich K, Peitsch MC and Hoeng J (2019) Construction of a Suite of Computable Biological Network Models Focused on Mucociliary Clearance in the Respiratory Tract. Front. Genet. 10:87. doi: 10.3389/fgene.2019.00087

INTRODUCTION

The respiratory tract is under constant challenge to provide the body with oxygen while monitoring air quality for pollutants and microorganisms. The mucous membranes in the airways, which are lined with microtubule-based projections, the cilia, represent a powerful first-line defense. In response to irritants and infection, mucus is secreted by goblet cells, and cilia on the surface of ciliated cells move mucus upward in coordinated waving and beating motions. Eventually, particles are expelled through sneeze and cough (Wanner et al., 1996). This self-clearing mechanism, mucociliary clearance (MCC), ensures proper functioning of the respiratory tract.

Cilia have attracted increasing attention because of the growing number of diseases caused by mutations in genes that impact cilium assembly, function, and turnover (Fliegauf et al., 2007; Kempeneers and Chilvers, 2018). Traditionally, cilia are classified as primary or motile (Wheatley, 1995; Satir and Christensen, 2007). Primary cilia are present on almost all cell types and are involved in tissue homeostasis (Gerdes et al., 2009;

Nozawa et al., 2013). Motile cilia often occur as clusters of several hundred protrusions covering cells and direct fluid flow (Choksi et al., 2014).

Cilia assembly and resorption often depend on the cell cycle (Kim and Tsiokas, 2011), with a neatly interwoven mode of regulation assuring timely and developmentally precise control of cilium biogenesis. The regulatory factor X (RFX) family of transcription factors is a key regulator of both primary and motile cilia assembly programs (reviewed in Thomas et al., 2010; Choksi et al., 2014). A master regulator of motile cilia assembly across the vertebrates is forkhead box J (FOXJ1), a member of the forkhead/winged-helix family of transcription factors (Murphy et al., 1997; Chen et al., 1998; Brody et al., 2000), which is under control of multiciliate differentiation and DNA synthesis-associated cell cycle protein (MCIDAS) and geminin coiled-coil domain-containing (GMNC) protein in the respiratory epithelium (Stubbs et al., 2012; Arbi et al., 2016). Mutations in MCIDAS and its downstream effector cyclin O are implicated in an MCC disorder known as reduced generation of multiple motile cilia (RGMC) (Boon et al., 2014). In RGMC patients, cilia numbers are reduced, resulting in impaired MCC, airway obstruction, and recurring respiratory infections.

An alternative mechanism of ciliary regulation is the disassembly of the organelle by aurora A kinase (AURKA), which also regulates the entry into mitosis (Pan et al., 2004; Pugacheva et al., 2007). AURKA phosphorylates histone deacetylase 6 (HDAC6), stimulating HDAC6-dependent deacetylation of axonemal microtubules (Hubbert et al., 2002), destabilization of the ciliary shaft, and subsequent collapse of the cilium.

Exposure to air pollutants, cigarette smoke, drugs, or infectious agents can affect ciliary beating frequency (CBF) (Workman and Cohen, 2014; Yaghi and Dolovich, 2016). On the molecular level, CBF increases in response to high mucus viscosity (Fernandes et al., 2008) and fluctuations in the levels of second messengers, such as cyclic adenosine 3′,5′-monophosphate (cAMP), cyclic guanidine 3′,5′-mono- phosphate (cGMP), intracellular Ca^{2+}, calmodulin, nitric oxide (Jain et al., 1993; Korngreen and Priel, 1996; Yang et al., 1996; Wyatt et al., 1998; Zagoory et al., 2001, 2002), and intracellular pH (Sutto et al., 2004). Mechanistically, CBF increases as a result of cAMP- and cGMP-mediated activation of respective protein kinases via Ca^{2+} release or by a calcium-independent mechanism.

While mucus secretion is a normal defense response, mucin synthesis in goblet cells and mucus secretion are amplified in respiratory diseases such as asthma or chronic obstructive pulmonary disease (COPD). In addition, the number of goblet cells can increase by proliferation (hyperplasia) and by airway epithelial cell transdifferentiation (metaplasia), further contributing to increased mucus production (Blyth et al., 1998; Rogers, 2007; Turner and Jones, 2009; Boucherat et al., 2013; Ramos et al., 2014). This airway epithelial remodeling decreases ciliated cell numbers and ciliary beating efficiency, reducing MCC and aggravating airway plugging (Nini et al., 2012; Yaghi et al., 2012; Yaghi and Dolovich, 2016).

There is overwhelming evidence that oxidative stress and oxidative damage play a pivotal role in the pathogenesis of COPD (Rahman and MacNee, 1999; Rahman and Adcock, 2006; Anderson and Macnee, 2009; Kim and Criner, 2015; Matera et al., 2016). Oxidative stress is a well-described trigger of the epidermal growth factor receptor (EGFR) signaling pathway that leads to mucus hypersecretion (Takeyama et al., 1999, 2001; Perrais et al., 2002; Hewson et al., 2004; Casalino-Matsuda et al., 2006; Hao et al., 2014). We recently published an adverse outcome pathway that describes the events that follow oxidative stress-mediated EGFR activation to goblet cell hyperplasia/metaplasia and decreased lung function following mucus overproduction (Luettich et al., 2017).

Signaling downstream of interleukin (IL) 13 is involved in the pathogenesis of asthma (Wills-Karp, 2004). The IL13 receptor complex initiates several cascades of molecular events that result in goblet cell metaplasia/hyperplasia. One important downstream effector of IL13 is the sterile alpha motif pointed domain-containing ETS transcription factor (SPDEF), which is directly involved in mucin gene expression (Park et al., 2007; Chen et al., 2009).

The vast volume and diversity of biological data available on cilium assembly, CBF, and goblet cell hyperplasia/metaplasia require that the information be integrated for better visualization and understanding of the processes that underlie respiratory diseases. Biological network models offer a framework for understanding biological processes and diseases and aid in drawing new, often unpredicted conclusions. Over the years, we have built several causal biological network (CBN) models that capture biological processes that are impacted in COPD. These models, stored in the CBN database, are emerging as an innovative and powerful tool to quantify the impact of exposure or potentially affected biological processes in disease (Cho et al., 2012; Martin et al., 2014; Boue et al., 2015; Talikka et al., 2017). The major advantage of the CBN approach is that it transforms unstructured data into interconnected and organized knowledge that describes biological processes precisely and accurately (Schlage et al., 2011; Westra et al., 2011, 2013; Cho et al., 2012; Gebel et al., 2013; De Leon et al., 2014; Martin et al., 2014; Szostak et al., 2016).

In this study, we present a suite of causal biological models that describe important molecular events involved in MCC, from cilium assembly to ciliary beating, goblet cell hyperplasia/metaplasia, and mucus hypersecretion. We also show how transcriptomic data are scored onto these network models and how the models can provide mechanistic understanding of gene expression changes.

MATERIALS AND METHODS

Literature Curation

Biological Expression Language (BEL)[1] version 1.0 is used for scientific text curation. BEL is a computable language that converts causal and correlative biological observations to statements consisting of two biological entities connected by a relationship predicate[1]. Relevant original research articles for

[1]http://openbel.org/

curation were identified from pertinent review articles in the field. The journal impact factor or any other means to rank the publications was not considered. If the statements in the original research articles were sufficiently supported by the results presented in figures, the information was considered reliable and captured. To retrieve causal relationships the result sections were extracted from these articles for curation. The introduction, discussion and conclusion sections were avoided because the evidences therein largely contain data from earlier studies, repetition of the results, hypotheses and assumptions. Although several evidences supporting an interaction would provide more confidence on the edge, we capture the interactions even when a single experiment is provided in the literature, in order to not omit the relevant information. Contradicting statements were captured without preferential treatment and with proper annotations (model organism, tissue, cell line, treatment/disease, experimental setup). The experimental information from the relevant peer-reviewed scientific articles is semi-automatically processed through the BEL Information Extraction workFlow (BELIEF) platform (Szostak et al., 2015, 2016). BELIEF contains a text-mining software that recognizes biological terms in the text and assembles them into BEL statements. The curation interface allows review, correction, and annotations (cell/tissue type, disease if applicable, species, and experimental design) of the statements that BELIEF proposes. The literature curation is an iterative process. After the curation of initial articles, a gap analysis is performed, and more literature is identified based on gaps in the network models.

Network Model Assembly and Visualization

BEL statements are then compiled to generate a cohesive knowledge assembly model using the OpenBEL framework 3.0.0, an open source compilation framework. The network model consists of nodes that are the biological entities in the network models connected by edges (i.e., the relationships between the biological entities). Any RNA nodes are removed from the model backbone and used in the downstream layer for model scoring as described in Martin et al. (2014). The Cytoscape web application[2] is used to visualize and analyze the network properties (Shannon et al., 2003). Cytoscape supports powerful visual mapping whereby biological entities are depicted as defined-shaped nodes connected by the relationship edges. The network visualization is used also during the curation process to identify the gaps and to trim the network models. The trimming here means that any nodes that are "hanging" and do not lead to a biological process described in the model are removed, or further curation is performed to add molecular relationships to connect such nodes to the biological process.

The network model suite is available in the CBN database. The NPA algorithm as well as some measurable "downstream" relationships (backbone node to mRNA) can be downloaded as R packages from the GitHub project pages https://github.com/pmpsa-hpc/NPA and https://github.com/pmpsa-hpc/NPAModels.

[2]http://www.cytoscape.org/

Network Model Scoring

The network perturbation amplitude (NPA) methodology is used to obtain a quantitative assessment of how each of the models interprets the transcriptomic changes in the datasets we selected (GSE22430, GSE37693, and GSE5264). This methodology allows for the translation of gene expression fold-changes to differential values for each network node as well as enabling a network-level summary to provide a quantitation of the degree of network model perturbation (Hoeng et al., 2012; Martin et al., 2014; Sewer et al., 2015; Szostak et al., 2016). Raw data were obtained from Gene Expression Omnibus (GEO) repository and normalized following a standard pipeline based on robust multiarray normalization implemented in the R environment for statistical computing (Smyth, 2004). The differential expression values and statistics were calculated using the Bioconductor LIMMA package with appropriate experimental comparisons. "O" and "K" statistics was used to test the specificity of the network models (including the "downstream edges" that connect the network nodes to gene differential expression nodes according to the underlying reverse-causal concept; Catlett et al., 2013). They compare the actual NPA value to the distributions of alternative NPA values obtained by permuting the edges of the networks (the connections between nodes for "K" and the connection between nodes and gene differential expression nodes for "O"). If the actual NPA value is significantly different from these "background" non-biological values, then we consider it as significantly specific.

The leading node analysis allows to focus on a fewer number of nodes in the network by ranking the nodes based on their contribution (%). Using an empirical 80% collective contribution instead of the actual rank, does not limit the number of the nodes, when the contribution of several nodes is almost equal (Martin et al., 2014).

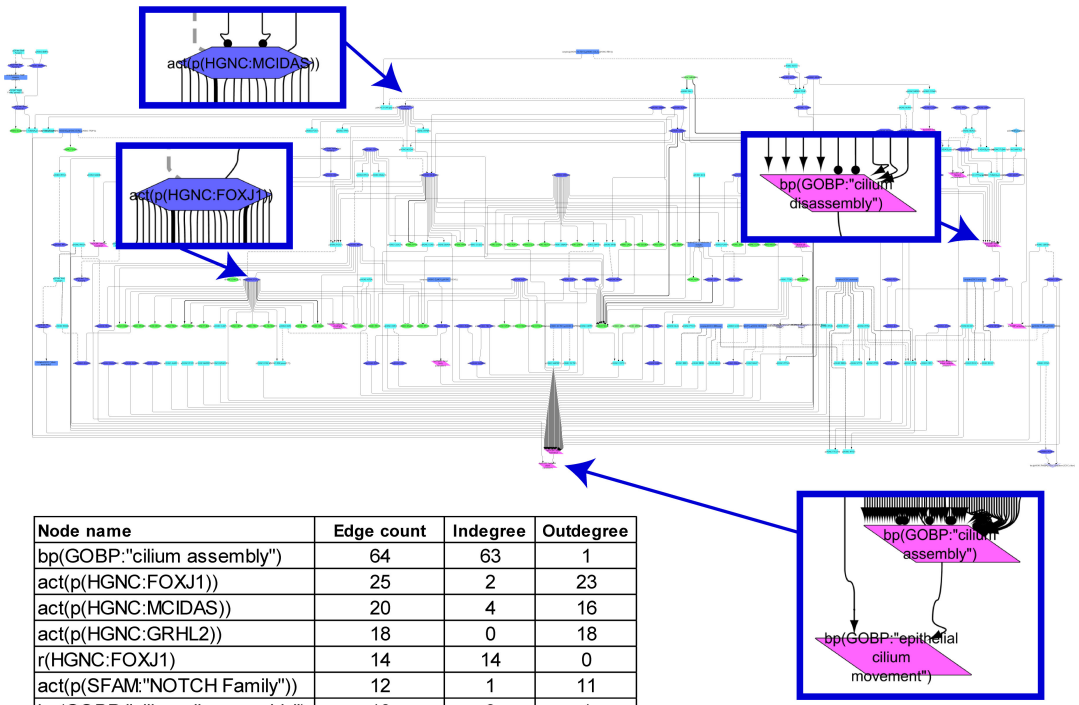
RESULTS

Model Description

Cilium Assembly Model

The cilium assembly network model is a collection of intertwined biological entities and processes that are supported by 59 relevant peer-reviewed articles. The network contains 209 nodes and 319 edges that represent relationships between nodes (**Figure 1**). When the connections between the nodes in the network were analyzed, many poorly connected nodes and a few highly connected ones, "hubs," were observed. The most connected node (63 indegree edges) was the biological process "cilium assembly," and the transcription factor FOXJ1, which is downstream of MCIDAS and GMNC, had the most outdegree edges (**Figure 1**).

A number of pathways, including delta-like canonical notch ligand (DLL)/NOTCH (through MCIDAS/GMNC), smoothened/hedgehog, and grainyhead-like transcription factor 2, converge into a FOXJ1/RFX module that triggers cilium assembly in the network model. This shows a high level of cooperativity between FOXJ1 and RFX factors; FOXJ1 can induce RFX2 and RFX3 expression, FOXJ1 gene expression is partially dependent on RFX3 activity, and a subset of FOXJ1



Node name	Edge count	Indegree	Outdegree
bp(GOBP:"cilium assembly")	64	63	1
act(p(HGNC:FOXJ1))	25	2	23
act(p(HGNC:MCIDAS))	20	4	16
act(p(HGNC:GRHL2))	18	0	18
r(HGNC:FOXJ1)	14	14	0
act(p(SFAM:"NOTCH Family"))	12	1	11
bp(GOBP:"cilium disassembly")	10	9	1
complex(GOCC:axoneme)	10	0	10
act(p(HGNC:GMNC))	10	3	7
act(p(HGNC:RFX3))	9	1	8

FIGURE 1 | Causal biological network model for cilium assembly. The table shows the top 10 highly connected nodes and their degrees of distribution. The vocabulary for the BEL is provided in http://www.openbel.org/. The Cytoscape layout is the Yfiles hierarchical layout. The network model can be downloaded from causalbionet.com.

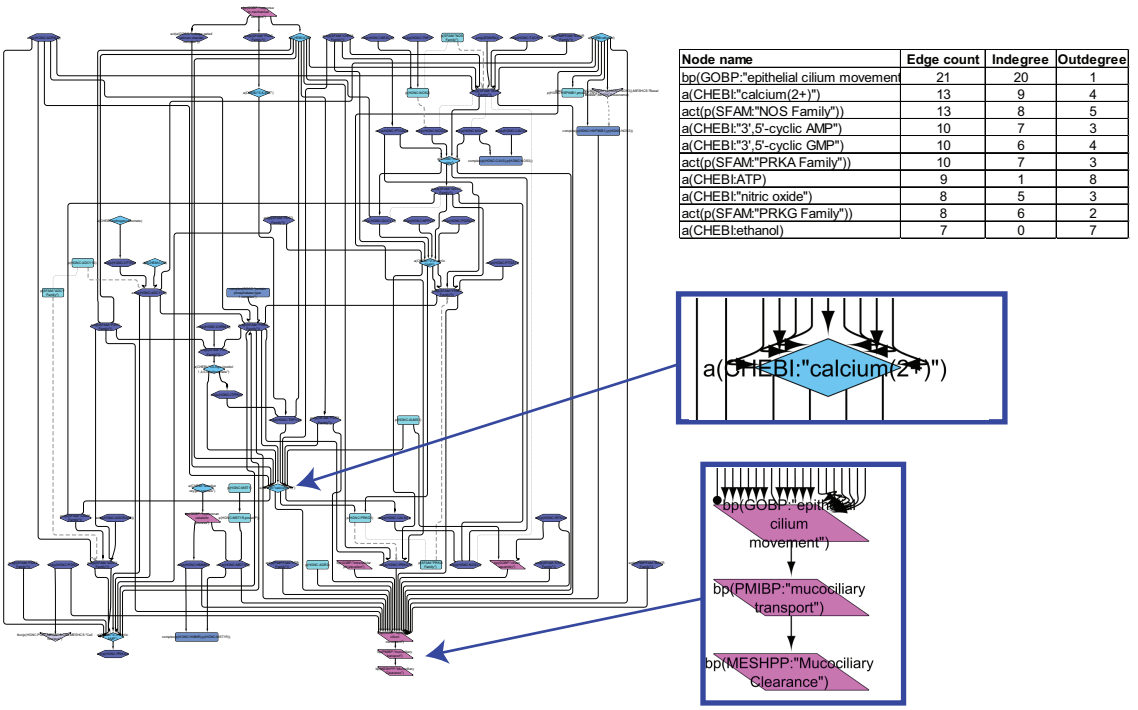
and RFX target genes overlap (**Figure 1**). This assures timely and developmentally precise control of cilium biogenesis. In addition, numerous molecules and complexes necessary for structural integrity of cilia, such as the axoneme constituents, BBSome complex (structural components of the basal body), and exocyst complex (membrane transport to cilium), support the "cilium assembly" hub as immediate neighbor nodes. With regard to the "cilium disassembly" hub, as expected, the AURKA-HDAC6 axis and their upstream regulators emerged as a supporting subnetwork.

Ciliary Beating Model

The ciliary beating network model was computed from 52 articles and comprises 80 nodes and 137 edges. The network illustrates the path from various stimuli through intermediate signaling molecules converging into consecutive biological processes, with "mucociliary clearance" as the final node (**Figure 2**). "Epithelial cilium movement" has the most inward connections in the network, and adenosine triphosphate has the most outward connections. Calcium and "nitric oxide synthase family" are central hubs in the network, with several incoming and outgoing edges. The model shows the CBF increases as a result of cAMP- and cGMP-mediated activation of the respective protein kinases through Ca^{2+} release or by a calcium-independent mechanism. The model also captures cystic fibrosis transmembrane conductance regulator, whose activation triggers the adenylate cyclase (ADCY)/cAMP pathway. Several other stimuli, such as serotonin or macrophage-stimulating protein, via corresponding receptors (HTR and MST1R, respectively), lead to increased ciliary motion in the model. Another level of regulation is added through sex hormone-dependent modulation, such as progesterone-mediated decreases or estrogen-mediated increases in CBF.

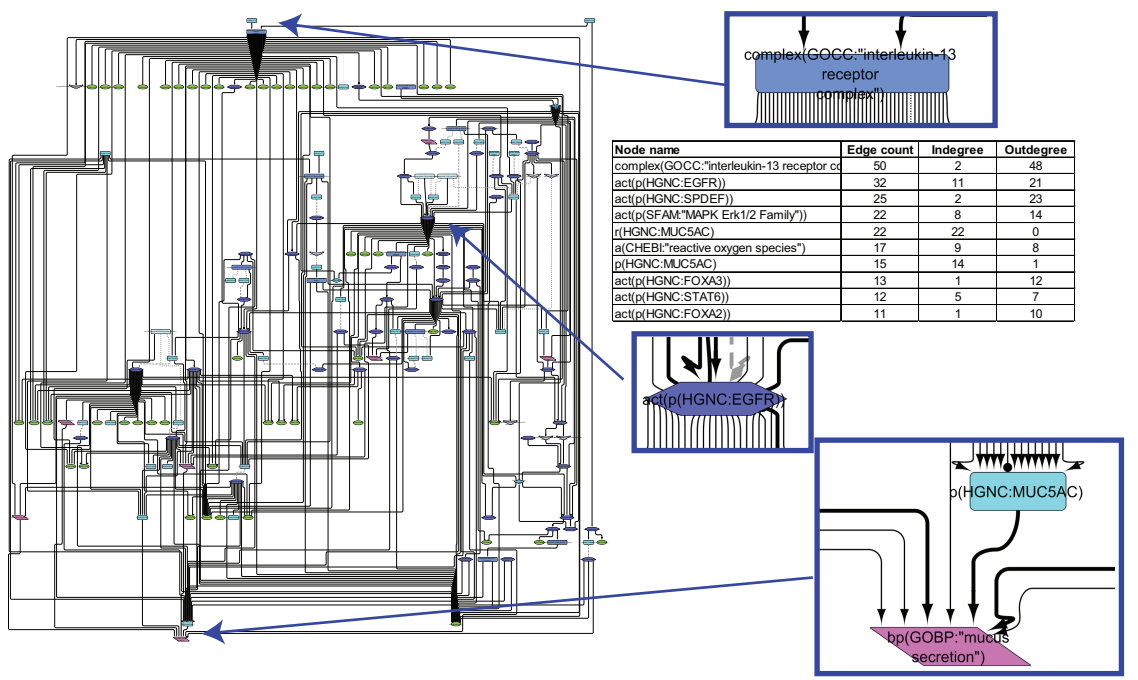
Goblet Cell Hyperplasia/Metaplasia Model

The goblet cell hyperplasia/metaplasia model covers 172 nodes and 335 edges that were obtained from 58 articles. The hierarchical view of the network model clearly indicates that, as expected, the biological process "mucus secretion" is the endpoint of the model (**Figure 3**). The network model hinges on EGFR and IL signaling pathways (**Figure 3**). An array of growth factors such as epidermal growth factor, transforming growth factor, tumor necrosis factor, amphiregulin, IL4, IL6, IL7, IL8, and IL13 initiate goblet cell-specific mucus secretion by activating their respective receptors (EGFR and IL6R, IL13R, IL17R) and subsequent signaling events, notably through Ras/Raf/mitogen-activated protein kinase kinase/mitogen-activated protein kinase (MAPK)/extracellular signal-regulated kinase 1/2 (ERK1/2) or janus kinase/signal transducer and activator of transcription/SPDEF effectors, modulating mucin



Node name	Edge count	Indegree	Outdegree
bp(GOBP:"epithelial cilium movement	21	20	1
a(CHEBI:"calcium(2+)")	13	9	4
act(p(SFAM:"NOS Family"))	13	8	5
a(CHEBI:"3',5'-cyclic AMP")	10	7	3
a(CHEBI:"3',5'-cyclic GMP")	10	6	4
act(p(SFAM:"PRKA Family"))	10	7	3
a(CHEBI:ATP)	9	1	8
a(CHEBI:"nitric oxide")	8	5	3
act(p(SFAM:"PRKG Family"))	8	6	2
a(CHEBI:ethanol)	7	0	7

FIGURE 2 | Causal biological network model for ciliary beating. The table shows the top 10 highly connected nodes and their degrees of distribution. The vocabulary for the BEL is provided in http://www.openbel.org/. The Cytoscape layout is the Yfiles hierarchical layout. The network model can be downloaded from causalbionet.com.



Node name	Edge count	Indegree	Outdegree
complex(GOCC:"interleukin-13 receptor co	50	2	48
act(p(HGNC:EGFR))	32	11	21
act(p(HGNC:SPDEF))	25	2	23
act(p(SFAM:"MAPK Erk1/2 Family"))	22	8	14
r(HGNC:MUC5AC)	22	22	0
a(CHEBI:"reactive oxygen species")	17	9	8
p(HGNC:MUC5AC)	15	14	1
act(p(HGNC:FOXA3))	13	1	12
act(p(HGNC:STAT6))	12	5	7
act(p(HGNC:FOXA2))	11	1	10

FIGURE 3 | Causal biological network model for goblet cell hyperplasia/metaplasia. The table shows the top 10 highly connected nodes and their degrees of distribution. The vocabulary for the BEL is provided in http://www.openbel.org/. The Cytoscape layout is the Yfiles hierarchical layout. The network model can be downloaded from causalbionet.com.

gene expression. Multiple additional factors leading to mucus hypersecretion and their interactions are also depicted in the network model. FOXA2 transcription factor, in contrast, limits goblet cell differentiation in the lung and directly represses mucin gene expression. The network displays the inhibition of FOXA2 by EGFR and IL13 pathways that results in goblet cell hyperplasia and mucus secretion.

Model Scoring With Transcriptomic Data

NPA

Network scoring with transcriptomic data is based on the inference of activities of the molecular entities in the network from gene expression changes. This backward reasoning employs a downstream layer with information on gene expression changes known to be induced by the backbone entities (Martin et al., 2014). To test the ability of the MCC network models to provide a quantitative measure of MCC, we identified publicly available datasets in Gene Expression Omnibus. The first dataset selected for model scoring (GSE22430) was from lung epithelial cells treated with the redox-active toxin pyocyanin from *Pseudomonas aeruginosa* that stimulates EGFR (Rada et al., 2011). Dataset GSE5264 was derived from an *in vitro* experiment, in which airway epithelial cells were allowed to differentiate to a pseudostratified epithelium at the air-liquid interface (Ross et al., 2007). Finally, we used a transcriptomic dataset from IL13-treated human airway epithelial cells (GSE37693) (Alevy et al., 2012).

The cilium assembly network model responded strongly to the treatment of lung cells with pyocyanin and to the time-course of bronchial epithelial cell differentiation with increasing amplitude over time. There was no impact on the models in response to the IL13 treatment (**Figure 4**). When the same datasets were used to score the cilia beating network models, the largest amplitude of network perturbation was observed in response to pyocyanin treatment of lung cells (**Figure 4**). Similar to the cilium assembly model, the amplitude of cilia beating network perturbation increased with advanced mucociliary differentiation, and the model did not respond to the IL13 treatment. The scoring of the goblet cell hyperplasia/metaplasia network model again showed a very strong response to the pyocyanin treatment and, to a lesser extent, to mucociliary differentiation of airway cells. This model responded to IL13 treatment (**Figure 4**).

Leading Node Analysis

To investigate the mechanistic foundation underlying the perturbations of the network models from transcriptomic data and to further validate the biology in the models, we used the leading node analysis (Martin et al., 2014). Leading nodes are the entities in the network models upon which the impact contributes 80% of the observed effect on the network as a whole. Leading node analysis also allows for the assessment of the directionality (activation or inhibition) of the inferred effect on each node. All leading nodes for all contrasts and models are provided in **Supplementary Data Sheets S1–S3**.

Cilium assembly model

Figure 5 shows the leading node analysis of the cilium assembly network model scored with transcriptomic data from early, intermediate, and late time points of human airway epithelial cell mucociliary differentiation. At the early time point, bone morphogenic protein (BMP) signaling was inferred to be upregulated. The mechanistic target of rapamycin (mTOR), platelet-derived growth factor A (PDGFA), and protein kinase B (AKT) signaling were inferred to be downregulated, in contrast with the inferred upregulation of cilium assembly. At the same time, NudE neurodevelopment protein 1 like 1 (NDEL1) was inferred to be downregulated, resulting in downregulation of cyclin A2 (CCNA2) and cell cycle arrest. At the intermediate and late time points of mucociliary differentiation, BMP signaling was no longer inferred to be upregulated. Instead, DLL1/NOTCH1 signaling was inferred to be downregulated, resulting in an increase in MCIDAS and FOXJ1, the master transcription factors required for the formation of motile cilia. RFX3, also known to induce FOXJ1, was inferred to be upregulated at the intermediate

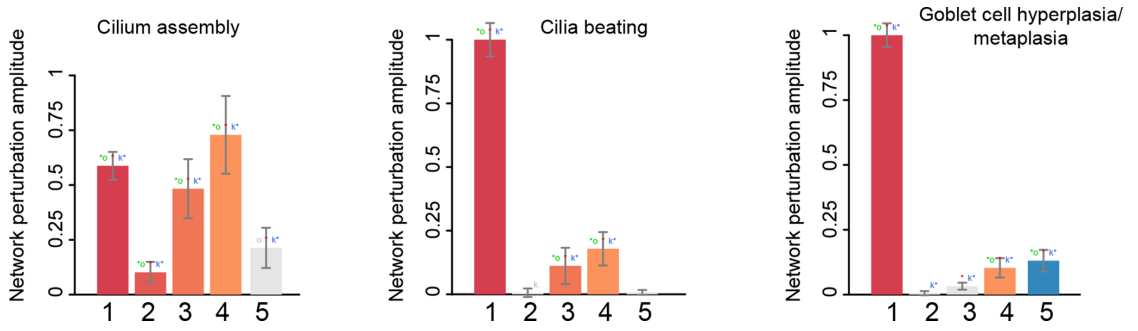


FIGURE 4 | The network perturbation amplitude (NPA). The NPA scores are shown with their confidence interval, accounting for experimental variation. The red star indicates that NPA is statistically different from 0. In addition, companion statistics derived to inform on the specificity of the NPA score with respect to the network structure are shown as *O and K*, respectively, if their *p*-values are below the significance level of 0.05, and by O and K when the corresponding *p*-values are between 0.05 and 0.1. * indicates O and K statistic *p*-values below 0.05 (in color), O and K *p*-values between 0.05 and 0.1 (in gray). Lanes: 1. GSE22430, pyocyanin-treated vs. control; 2. GSE5264, early time points of mucociliary differentiation in human airway epithelial cells; 3. GSE5264, intermediate time points of mucociliary differentiation in human airway epithelial cells; 4. GSE5264, late time points of mucociliary differentiation in human airway epithelial cells; 5. GSE37693, IL13-treated vs. control.

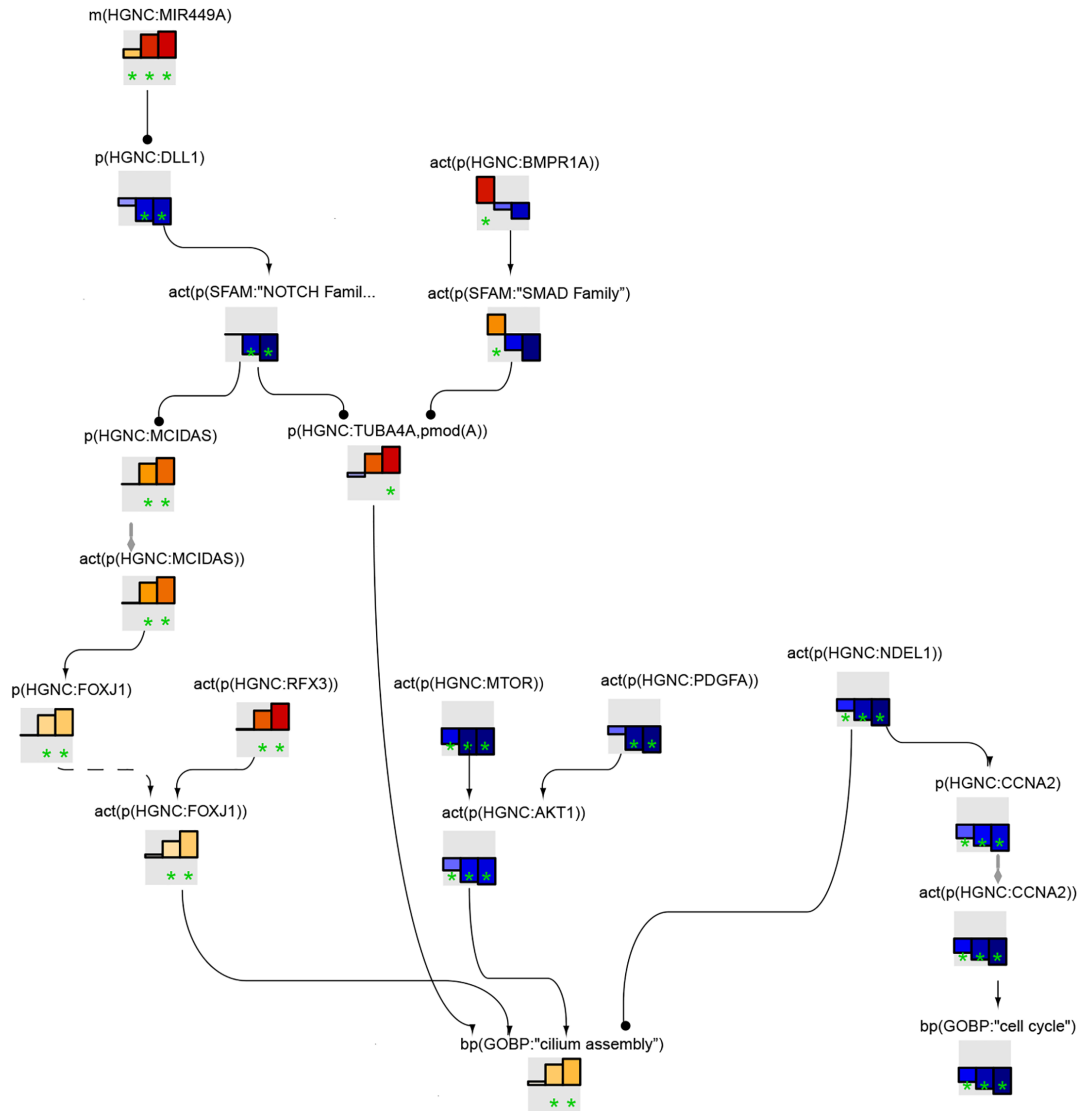


FIGURE 5 | Cilium assembly subnetwork based on leading nodes. The graphical representation shows an example of connected leading nodes common for early (first lane), intermediate (second lane), and late (third lane) time points of human airway epithelial cell mucociliary differentiation. The top 10 leading nodes were prioritized and connected with other leading nodes from the analysis. The backbone NPA values with directionalities of inferred regulation are shown as bar graphs for each node. Orange/red bars indicate inferred upregulation and blue bars indicate inferred downregulation. The green asterisk indicates that the node is a leading node. The vocabulary for the BEL is provided in http://www.openbel.org/.

and late time points. PDGFA, mTOR, AKT, and NDEL1/CCNA2 continued to be downregulated in the leading node analysis.

The leading node analysis of the pyocyanin treatment data scored on the cilium assembly network model indicated the upregulation of PDGFA and downregulation of BMP signaling (**Supplementary Data Sheet S1**).

Ciliary beating model

Figure 6 shows the leading node analysis of the ciliary beating network model scored with transcriptomic data from early, intermediate, and late time points of human airway epithelial cell mucociliary differentiation. The β2-adrenergic receptor/ADCY signaling pathway, leading to an increase in cAMP levels and subsequent Ca^{2+} increase via the activation of the PRKA family, was inferred to be upregulated. The analysis also inferred the activation of cGMP-dependent protein kinase 1 (PRKG1). In addition, the leading node analysis of the pyocyanin treatment data scored on the ciliary beating network model indicated the upregulation of ADCY and calcium signaling (**Supplementary Data Sheet S2**).

Goblet cell hyperplasia/metaplasia model

Figure 7 shows the leading-node analysis of the goblet cell hyperplasia/metaplasia model with the pyocyanin and IL13 datasets. The levels of reactive oxygen species (ROS) were inferred to increase with subsequent activation of EGFR

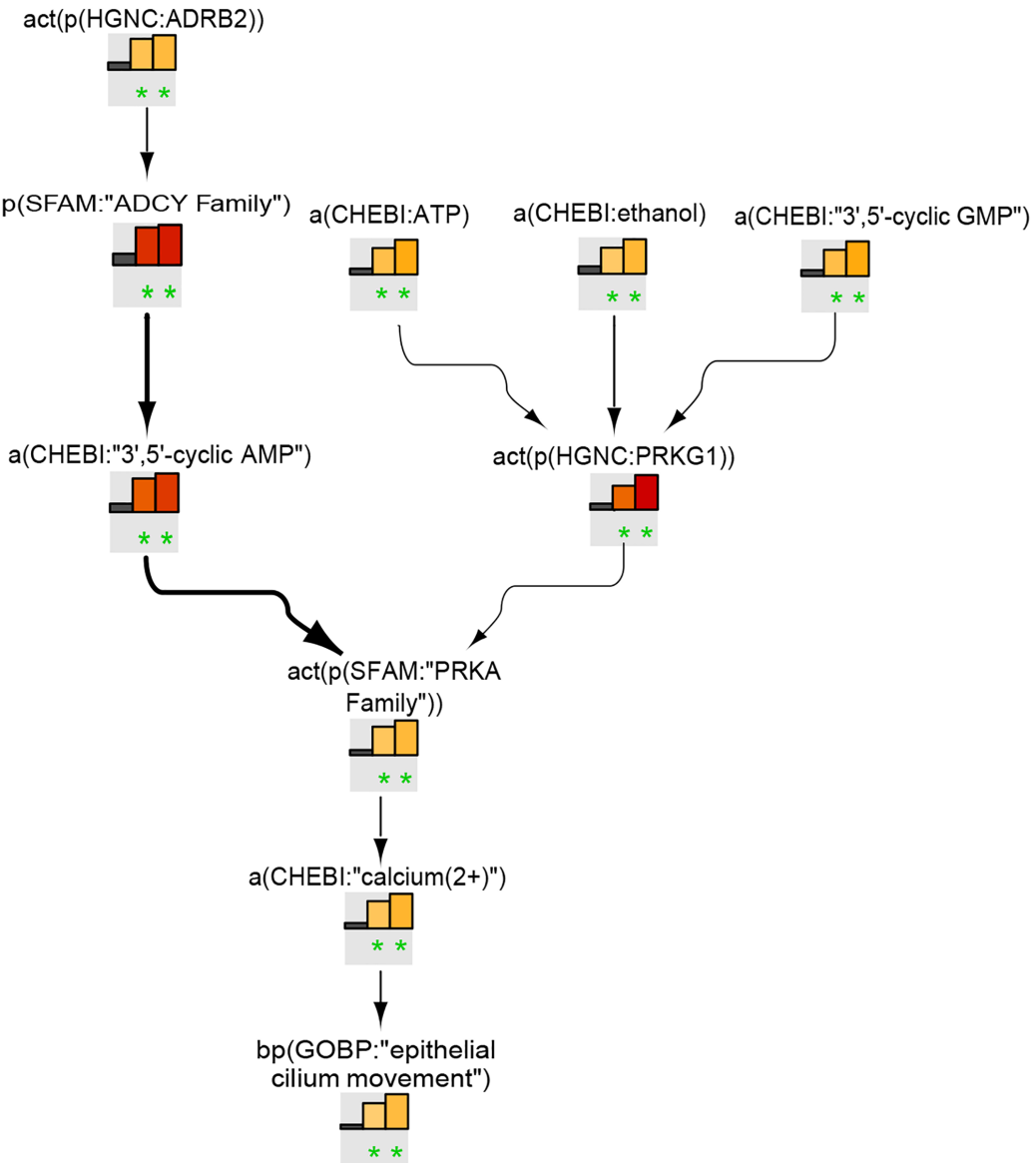


FIGURE 6 | Ciliary beating subnetwork based on leading nodes. The graphical representation shows an example of connected leading nodes common for early (first lane), intermediate (second lane), and late (third lane) time points of human airway epithelial cell mucociliary differentiation. The top 10 leading nodes were prioritized and connected with other leadin g nodes from the analysis. The backbone NPA values with directionalities of inferred regulation are shown as bar graphs for each node. Orange/red bars indicate inferred upregulation and blue bars indicate inferred downregulation. Bold edges indicate "direct regulation." The green asterisk indicates that the node is a leading node. The vocabulary for the BEL is provided in http://www.openbel.org/.

and ERK1/2, followed by an inferred increase in mucin production. Three other branches of the network that were highlighted and led to increases in mucins included the AP-1, FOXA3/SPDEF, and IL13/SPDEF pathways. The inferred activation of the IL13 receptor complex mirrored the activation of the matrix metalloproteinase family. The inferred activation of the p38 MAPK family upstream of MUC5AC was unique to the pyocyanin dataset. While the NPA analysis showed a significant network perturbation in response to human airway epithelial cell mucociliary differentiation, a closer examination of the leading nodes clearly indicated inferred downregulation of ROS and the EGFR and MAPK ERK1/2 pathways (**Supplementary Data Sheet S3**).

DISCUSSION

MCC is an important defense mechanism that protects the respiratory tract, and thus the body, from infections and airborne pollutants. In this article, we presented a suite of CBN models that describe relevant molecular processes related to MCC. Derived from original articles, the BEL-scripted scientific statements

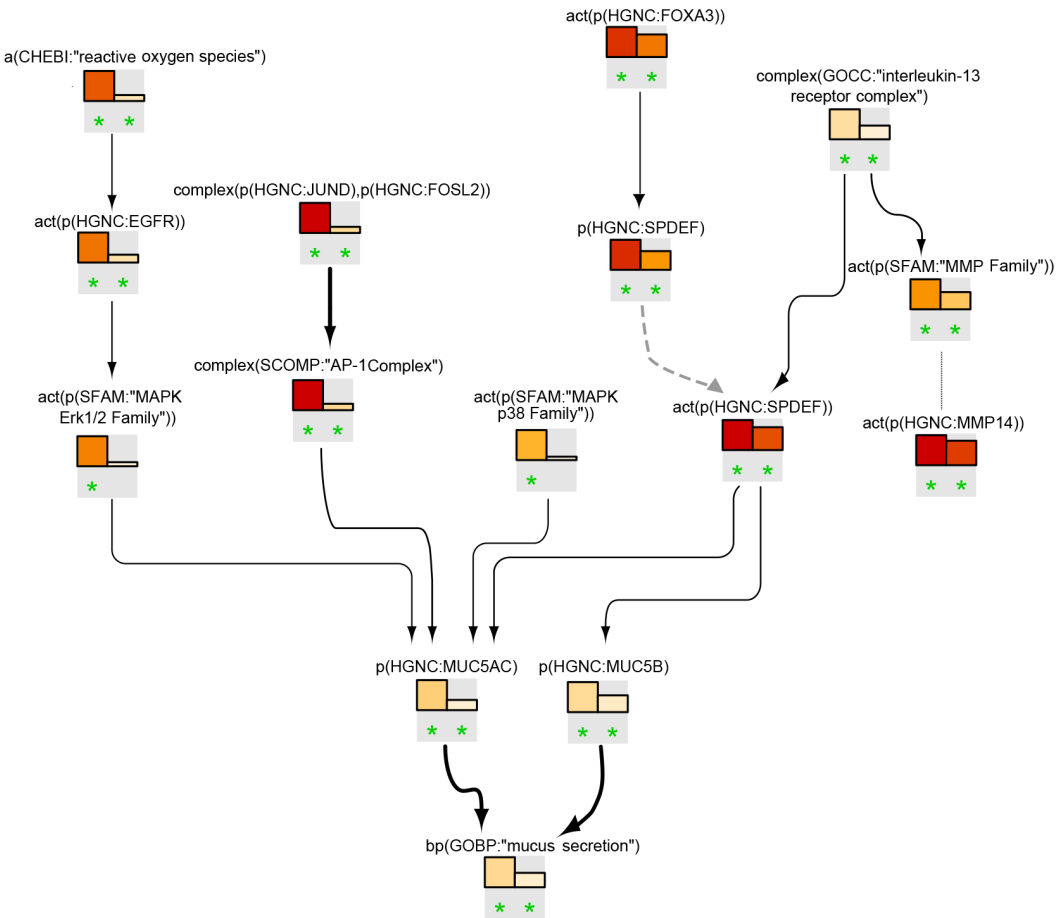


FIGURE 7 | Goblet cell hyperplasia/metaplasia subnetwork based on leading nodes. The graphical representation shows an example of connected leading nodes for pyocyanin treatment (first lane) and IL13 treatment (second lane) compared with experimental controls. The top 10 leading nodes were prioritized and connected with other leading nodes from the analysis. Orange/red bars indicate inferred upregulation. Bold edges indicate "direct regulation." The dotted edge denotes a member of a protein family. The green asterisk indicates that the node is a leading node in a given contrast. The vocabulary for the BEL is provided in http://www.openbel.org/.

were assembled into three separate network models that capture molecular processes involved in cilium assembly, ciliary beating, and goblet cell hyperplasia/metaplasia accurately. The key factors involved in these processes are part of the backbone that interconnects various entities in the network models. As an example, the cilium assembly hub integrates the diversity of the cascades that are determined by the variety of cilia types, each requiring precise regulation (for review, see Choksi et al., 2014).

As part of network model validation, we conducted network scoring with gene expression data from experiments that were expected to trigger perturbation of the MCC models. The scoring also allowed us to look farther from the static network view into the key factors that impact the network and assess the behavior (activation or inhibition) of molecular entities in the model backbone based on differential gene expression in the selected datasets.

As expected, the biology in the redox-active pyocyanin treatment experiment was best reflected in the goblet cell hyperplasia/metaplasia model, with EGFR and downstream MAPK ERK1/2 factors predicted to be activated, leading to mucin production. This was in line with other experimental observations of increased numbers of goblet cells and increased mucin production in response to pyocyanin treatment (Rada et al., 2011).

Impact on the cilium-focused network models could be explained by the cell redox state and ROS levels affecting multiple cellular signaling pathways, some of which overlap with cilium biology. As an example, the activity of the nitric oxide synthase (NOS) family as well as the nitric oxide (NO) chemical node were inferred to be upregulated by pyocyanin treatment in the ciliary beating network model (**Supplementary Data Sheet S2**). NO is a redox molecule that regulates tissue oxidative balance through direct and indirect mechanisms of action and can lead to an increase in ciliary beat frequency through the activation of NOS family.

Network scoring with the airway epithelial cell differentiation dataset clearly showed time-dependent activation of pathways leading to cilium assembly and ciliary beating. At the early stage, BMP signaling was inferred to be upregulated, indicating the lack of cilium assembly, while at later time points, BMP

released the inhibitory effect on cilium assembly, with the MCIDAS/FOXJ1/RFX3 pathway inferred to be activated to promote cilium assembly. Network scoring, however, indicated that mTOR and AKT were downregulated in the dataset, contradictory to the causal connection from mTOR and AKT to cilium assembly that was inferred to be upregulated. These relationships were derived from articles describing primary cilium assembly and may not appropriately reflect the biology in the respiratory tract, where the operating process is the motile cilia assembly program that culminates in the FOXJ1/RFX3 module (Wang et al., 2015; Suizu et al., 2016).

Downregulation of CCNA2 and cell cycle arrest in the cilium assembly network model could indicate a slowing down in cell proliferation to enforce cell differentiation. This was in accordance with the inferred inhibition of the mTOR/AKT pathway in the cilium assembly model. The results were further enforced by the scoring of the goblet cell network. The inferred reduction in EGFR signaling could indicate loss of proliferative potential in cultures differentiating to pseudostratified epithelium. This result also suggests that the network model discriminates between a physiological (i.e., differentiation) and pathological (i.e., COPD-related) increase in the number of goblet cells in airways. Finally, the cilia beating model appropriately captured the activation of cAMP/PRKA and cGMP/PRKG signaling that elevates cellular Ca^{2+} levels, leading to increases in cilia beating.

Scoring the three network models with the datasets from IL13-treated lung cells highlighted the specificities of the different networks: IL13-induced airway mucus production affected several hubs in the hyperplasia/metaplasia model, notably the SPDEF transcription factor, while impacts on the cilium assembly and ciliary beating models did not reach statistical significance.

In conclusion, the representation of cilium assembly, ciliary beating, and airway remodeling processes through CBN models is a potential powerful tool for systems medicine (Talikka et al., 2017). MCC networks can be used as a substrate for scoring high-throughput data for mechanistic understanding of the differences between diseased and healthy tissue. The MCC network model suite presented here, along with gene expression data from well-controlled clinical studies, could be used in individuals with MCC disorders for subject classification, identification of mode of action of novel drug candidates, or prediction of treatment outcome. Ultimately, the MCC network model suite provides perspectives for tailored drug therapy and precision medicine.

AUTHOR CONTRIBUTIONS

MT, JH, and MP conceived and designed the experiments. HY, SV, and MT performed the study. HY and MT wrote the manuscript with the support of KL. FM, AS, and SG analyzed the data. All authors made critical revision and approved the final version of the manuscript.

FUNDING

Philip Morris International is the sole source of funding and sponsor of this research.

ACKNOWLEDGMENTS

We would like to thank Nicholas Karoglou and Elena Scotti for editorial assistance as well as Stephanie Boue for preparing and uploading the network models in the CBN database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00087/full#supplementary-material

DATA SHEET S1 | Leading node analysis for cilium assembly network model scored with transcriptomic data from pyocyanin treatment, IL-13 treatment, and mucociliary differentiation datasets. For each node, an * indicates it is a leading node and the number denotes its rank among the leading nodes. + or − in parenthesis indicates inferred up- or down-regulation, respectively. Value in % is the contribution of the node to the overall NPA.

DATA SHEET S2 | Leading node analysis for ciliary beating network model scored with transcriptomic data from pyocyanin treatment, IL-13 treatment, and mucociliary differentiation datasets. For each node, an * indicates it is a leading node and the number denotes its rank among the leading nodes. + or − in parenthesis indicates inferred up- or down-regulation, respectively. Value in % is the contribution of the node to the overall NPA.

DATA SHEET S3 | Leading node analysis for goblet cell hyperplasia/metaplasia network model scored with transcriptomic data from pyocyanin treatment, IL-13 treatment, and mucociliary differentiation datasets. For each node, an * indicates it is a leading node and the number denotes its rank among the leading nodes. + or − in parenthesis indicates inferred up- or down-regulation, respectively. Value in % is the contribution of the node to the overall NPA.

REFERENCES

Alevy, Y. G., Patel, A. C., Romero, A. G., Patel, D. A., Tucker, J., Roswit, W. T., et al. (2012). IL-13–induced airway mucus production is attenuated by MAPK13 inhibition. *J. Clin. Invest.* 122, 4555–4568. doi: 10.1172/JCI64896

Anderson, D., and Macnee, W. (2009). Targeted treatment in COPD: a multi-system approach for a multi-system disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* 4, 321–335. doi: 10.2147/COPD.S2999

Arbi, M., Pefani, D. E., Kyrousi, C., Lalioti, M. E., Kalogeropoulou, A., Papanastasiou, A. D., et al. (2016). GemC1 controls multiciliogenesis in the airway epithelium. *EMBO Rep.* 17, 400–413. doi: 10.15252/embr.201540882

Blyth, D. I., Pedrick, M. S., Savage, T. J., Bright, H., Beesley, J. E., and Sanjar, S. (1998). Induction, duration, and resolution of airway goblet cell hyperplasia in a murine model of atopic asthma: effect of concurrent infection with respiratory syncytial virus and response to dexamethasone. *Am. J. Respir. Cell Mol. Biol.* 19, 38–54. doi: 10.1165/ajrcmb.19.1.2930

Boon, M., Wallmeier, J., Ma, L., Loges, N. T., Jaspers, M., Olbrich, H., et al. (2014). MCIDAS mutations result in a mucociliary clearance disorder with reduced generation of multiple motile cilia. *Nat. Commun.* 5:4418. doi: 10.1038/ncomms5418

Boucherat, O., Boczkowski, J., Jeannotte, L., and Delacourt, C. (2013). Cellular and molecular mechanisms of goblet cell metaplasia in the respiratory airways. *Exp. Lung Res.* 39, 207–216. doi: 10.3109/01902148.2013.791733

Boue, S., Talikka, M., Westra, J. W., Hayes, W., Di Fabio, A., Park, J., et al. (2015). Causal biological network database: a comprehensive platform of causal

biological network models focused on the pulmonary and vascular systems. *Database* 2015:bav030. doi: 10.1093/database/bav030

Brody, S. L., Yan, X. H., Wuerffel, M. K., Song, S. K., and Shapiro, S. D. (2000). Ciliogenesis and left-right axis defects in forkhead factor HFH-4-null mice. *Am. J. Respir. Cell Mol. Biol.* 23, 45–51. doi: 10.1165/ajrcmb.23.1.4070

Casalino-Matsuda, S. M., Monzon, M. E., and Forteza, R. M. (2006). Epidermal growth factor receptor activation by epidermal growth factor mediates oxidant-induced goblet cell metaplasia in human airway epithelium. *Am. J. Respir. Cell Mol. Biol.* 34, 581–591. doi: 10.1165/rcmb.2005-0386OC

Catlett, N. L., Bargnesi, A. J., Ungerer, S., Seagaran, T., Ladd, W., Elliston, K. O., et al. (2013). Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics* 14:340. doi: 10.1186/1471-2105-14-340

Chen, G., Korfhagen, T. R., Xu, Y., Kitzmiller, J., Wert, S. E., Maeda, Y., et al. (2009). SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J. Clin. Invest.* 119, 2914–2924. doi: 10.1172/JCI39731

Chen, J., Knowles, H. J., Hebert, J. L., and Hackett, B. P. (1998). Mutation of the mouse hepatocyte nuclear factor/forkhead homologue 4 gene results in an absence of cilia and random left-right asymmetry. *J. Clin. Invest.* 102, 1077–1082. doi: 10.1172/JCI4786

Cho, D. Y., Kim, Y. A., and Przytycka, T. M. (2012). Chapter 5: network biology approach to complex diseases. *PLoS Comput. Biol.* 8:e1002820. doi: 10.1371/journal.pcbi.1002820

Choksi, S. P., Lauter, G., Swoboda, P., and Roy, S. (2014). Switching on cilia: transcriptional networks regulating ciliogenesis. *Development* 141, 1427–1441. doi: 10.1242/dev.074666

De Leon, H., Boue, S., Schlage, W. K., Boukharov, N., Westra, J. W., Gebel, S., et al. (2014). A vascular biology network model focused on inflammatory processes to investigate atherogenesis and plaque instability. *J. Transl. Med.* 12:185. doi: 10.1186/1479-5876-12-185

Fernandes, J., Lorenzo, I. M., Andrade, Y. N., Garcia-Elias, A., Serra, S. A., Fernandez-Fernandez, J. M., et al. (2008). IP3 sensitizes TRPV4 channel to the mechano- and osmotransducing messenger 5′-6′-epoxyeicosatrienoic acid. *J. Cell Biol.* 181, 143–155. doi: 10.1083/jcb.200712058

Fliegauf, M., Benzing, T., and Omran, H. (2007). When cilia go bad: cilia defects and ciliopathies. *Nat. Rev. Mol. Cell Biol.* 8, 880–893. doi: 10.1038/nrm2278

Gebel, S., Lichtner, R. B., Frushour, B., Schlage, W. K., Hoang, V., Talikka, M., et al. (2013). Construction of a computable network model for DNA damage, autophagy, cell death, and senescence. *Bioinform. Biol. Insights* 7, 97–117. doi: 10.4137/BBI.S11154

Gerdes, J. M., Davis, E. E., and Katsanis, N. (2009). The vertebrate primary cilium in development, homeostasis, and disease. *Cell* 137, 32–45. doi: 10.1016/j.cell.2009.03.023

Hao, Y., Kuang, Z., Jing, J., Miao, J., Mei, L. Y., Lee, R. J., et al. (2014). Mycoplasma pneumoniae modulates STAT3-STAT6/EGFR-FOXA2 signaling to induce overexpression of airway mucins. *Infect. Immun.* 82, 5246–5255. doi: 10.1128/IAI.01989-14

Hewson, C. A., Edbrooke, M. R., and Johnston, S. L. (2004). PMA induces the MUC5AC respiratory mucin in human bronchial epithelial cells, via PKC, EGF/TGF-alpha, Ras/Raf, MEK, ERK and Sp1-dependent mechanisms. *J. Mol. Biol.* 344, 683–695. doi: 10.1016/j.jmb.2004.09.059

Hoeng, J., Deehan, R., Pratt, D., Martin, F., Sewer, A., Thomson, T. M., et al. (2012). A network-based approach to quantifying the impact of biologically active substances. *Drug Discov. Today* 17, 413–418. doi: 10.1016/j.drudis.2011.11.008

Hubbert, C., Guardiola, A., Shao, R., Kawaguchi, Y., Ito, A., Nixon, A., et al. (2002). HDAC6 is a microtubule-associated deacetylase. *Nature* 417, 455–458. doi: 10.1038/417455a

Jain, B., Rubinstein, I., Robbins, R. A., Leise, K. L., and Sisson, J. H. (1993). Modulation of airway epithelial cell ciliary beat frequency by nitric oxide. *Biochem. Biophys. Res. Commun.* 191, 83–88. doi: 10.1006/bbrc.1993.1187

Kempeneers, C., and Chilvers, M. A. (2018). To beat, or not to beat, that is question! The spectrum of ciliopathies. *Pediatr. Pulmonol.* 53, 1122–1129. doi: 10.1002/ppul.24078

Kim, S., and Tsiokas, L. (2011). Cilia and cell cycle re-entry: more than a coincidence. *Cell Cycle* 10, 2683–2690. doi: 10.4161/cc.10.16.17009

Kim, V., and Criner, G. J. (2015). The chronic bronchitis phenotype in chronic obstructive pulmonary disease: features and implications. *Curr. Opin. Pulm. Med.* 21, 133–141. doi: 10.1097/MCP.0000000000000145

Korngreen, A., and Priel, Z. (1996). Purinergic stimulation of rabbit ciliated airway epithelia: control by multiple calcium sources. *J. Physiol.* 497(Pt 1), 53–66. doi: 10.1113/jphysiol.1996.sp021749

Luettich, K., Talikka, M., Lowe, F. J., Haswell, L. E., Park, J., Gaca, M. D., et al. (2017). The adverse outcome pathway for oxidative stress-mediated EGFR activation leading to decreased lung function. *Appl. Vitro Toxicol.* 3, 99–109. doi: 10.1089/aivt.2016.0032

Martin, F., Sewer, A., Talikka, M., Xiang, Y., Hoeng, J., and Peitsch, M. C. (2014). Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC Bioinformatics* 15:238. doi: 10.1186/1471-2105-15-238

Matera, M. G., Calzetta, L., and Cazzola, M. (2016). Oxidation pathway and exacerbations in COPD: the role of NAC. *Expert Rev. Respir. Med.* 10, 89–97. doi: 10.1586/17476348.2016.1121105

Murphy, D. B., Seemann, S., Wiese, S., Kirschner, R., Grzeschik, K. H., and Thies, U. (1997). The human hepatocyte nuclear factor 3/fork head gene FKHL13: genomic structure and pattern of expression. *Genomics* 40, 462–469. doi: 10.1006/geno.1996.4587

Nini, G., Raica, M., Neamtiu, V., and Onel, M. (2012). Morphological study of bronchial mucosa in the chronic obstructive pulmonary disease under the influence of therapeutic algorithm. *Rom. J. Morphol. Embryol.* 53, 121–134.

Nozawa, Y. I., Lin, C., and Chuang, P. T. (2013). Hedgehog signaling from the primary cilium to the nucleus: an emerging picture of ciliary localization, trafficking and transduction. *Curr. Opin. Genet Dev.* 23, 429–437. doi: 10.1016/j.gde.2013.04.008

Pan, J., Wang, Q., and Snell, W. J. (2004). An aurora kinase is essential for flagellar disassembly in Chlamydomonas. *Dev. Cell* 6, 445–451. doi: 10.1016/S1534-5807(04)00064-4

Park, K. S., Korfhagen, T. R., Bruno, M. D., Kitzmiller, J. A., Wan, H., Wert, S. E., et al. (2007). SPDEF regulates goblet cell hyperplasia in the airway epithelium. *J. Clin. Invest.* 117, 978–988. doi: 10.1172/JCI29176

Perrais, M., Pigny, P., Copin, M. C., Aubert, J. P., and Van Seuningen, I. (2002). Induction of MUC2 and MUC5AC mucins by factors of the epidermal growth factor (EGF) family is mediated by EGF receptor/Ras/Raf/extracellular signal-regulated kinase cascade and Sp1. *J. Biol. Chem.* 277, 32258–32267. doi: 10.1074/jbc.M204862200

Pugacheva, E. N., Jablonski, S. A., Hartman, T. R., Henske, E. P., and Golemis, E. A. (2007). HEF1-dependent Aurora A activation induces disassembly of the primary cilium. *Cell* 129, 1351–1363. doi: 10.1016/j.cell.2007.04.035

Rada, B., Gardina, P., Myers, T. G., and Leto, T. L. (2011). Reactive oxygen species mediate inflammatory cytokine release and EGFR-dependent mucin secretion in airway epithelial cells exposed to *Pseudomonas* pyocyanin. *Mucosal Immunol.* 4, 158–171. doi: 10.1038/mi.2010.62

Rahman, I., and Adcock, I. M. (2006). Oxidative stress and redox regulation of lung inflammation in COPD. *Eur. Respir. J.* 28, 219–242. doi: 10.1183/09031936.06.00053805

Rahman, I., and MacNee, W. (1999). Lung glutathione and oxidative stress: implications in cigarette smoke-induced airway disease. *Am. J. Physiol.* 277, L1067–L1088. doi: 10.1152/ajplung.1999.277.6.L1067

Ramos, F. L., Krahnke, J. S., and Kim, V. (2014). Clinical issues of mucus accumulation in COPD. *Int. J. Chron. Obstruct. Pulmon. Dis.* 9, 139–150. doi: 10.2147/COPD.S38938

Rogers, D. F. (2007). Physiology of airway mucus secretion and pathophysiology of hypersecretion. *Respir. Care* 52, 1134–1146;discussion 1146–1139.

Ross, A. J., Dailey, L. A., Brighton, L. E., and Devlin, R. B. (2007). Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* 37, 169–185. doi: 10.1165/rcmb.2006-0466OC

Satir, P., and Christensen, S. T. (2007). Overview of structure and function of mammalian cilia. *Annu. Rev. Physiol.* 69, 377–400. doi: 10.1146/annurev.physiol.69.040705.141236

Schlage, W. K., Westra, J. W., Gebel, S., Catlett, N. L., Mathis, C., Frushour, B. P., et al. (2011). A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. *BMC Syst. Biol.* 5:168. doi: 10.1186/1752-0509-5-168

Sewer, A., Martin, F., Schlage, W. K., Hoeng, J., and Peitsch, M. C. (2015). "Quantifying the Biological Impact of Active Substances Using Causal Network Models," in *Computational Systems Toxicology*, eds J. Hoeng and M. Peitsch (New York, NY: Humana Press), 223–256.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet Mol. Biol.* 3:Article3. doi: 10.2202/1544-6115.1027

Stubbs, J. L., Vladar, E. K., Axelrod, J. D., and Kintner, C. (2012). Multicilin promotes centriole assembly and ciliogenesis during multiciliate cell differentiation. *Nat. Cell Biol.* 14, 140–147. doi: 10.1038/ncb2406

Suizu, F., Hirata, N., Kimura, K., Edamura, T., Tanaka, T., Ishigaki, S., et al. (2016). Phosphorylation-dependent Akt-Inversin interaction at the basal body of primary cilia. *EMBO J.* 35, 1346–1363. doi: 10.15252/embj.201593003

Sutto, Z., Conner, G. E., and Salathe, M. (2004). Regulation of human airway ciliary beat frequency by intracellular pH. *J. Physiol.* 560, 519–532. doi: 10.1113/jphysiol.2004.068171

Szostak, J., Ansari, S., Madan, S., Fluck, J., Talikka, M., Iskandar, A., et al. (2015). Construction of biological networks from unstructured information based on a semi-automated curation workflow. *Database* 2015:bav057. doi: 10.1093/database/bav057

Szostak, J., Martin, F., Talikka, M., Peitsch, M. C., and Hoeng, J. (2016). Semi-automated curation allows causal network model building for the quantification of age-dependent plaque progression in ApoE-/- mouse. *Gene Regul. Syst. Bio.* 10, 95–103. doi: 10.4137/GRSB.S40031

Takeyama, K., Dabbagh, K., Lee, H. M., Agusti, C., Lausier, J. A., Ueki, I. F., et al. (1999). Epidermal growth factor system regulates mucin production in airways. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3081–3086. doi: 10.1073/pnas.96.6.3081

Takeyama, K., Jung, B., Shim, J. J., Burgel, P. R., Dao-Pick, T., Ueki, I. F., et al. (2001). Activation of epidermal growth factor receptors is responsible for mucin synthesis induced by cigarette smoke. *Am. J. Physiol. Lung Cell Mol. Physiol.* 280, L165–L172. doi: 10.1152/ajplung.2001.280.1.L165

Talikka, M., Bukharov, N., Hayes, W. S., Hofmann-Apitius, M., Alexopoulos, L., Peitsch, M. C., et al. (2017). Novel approaches to develop community-built biological network models for potential drug discovery. *Expert Opin. Drug Discov.* 12, 849–857. doi: 10.1080/17460441.2017.1335302

Thomas, J., Morle, L., Soulavie, F., Laurencon, A., Sagnol, S., and Durand, B. (2010). Transcriptional control of genes involved in ciliogenesis: a first step in making cilia. *Biol. Cell* 102, 499–513. doi: 10.1042/BC20100035

Turner, J., and Jones, C. E. (2009). Regulation of mucin expression in respiratory diseases. *Biochem. Soc. Trans.* 37, 877–881. doi: 10.1042/BST0370877

Wang, S., Livingston, M. J., Su, Y., and Dong, Z. (2015). Reciprocal regulation of cilia and autophagy via the MTOR and proteasome pathways. *Autophagy* 11, 607–616. doi: 10.1080/15548627.2015.1023983

Wanner, A., Salathe, M., and O'Riordan, T. G. (1996). Mucociliary clearance in the airways. *Am. J. Respir. Crit. Care Med.* 154, 1868–1902. doi: 10.1164/ajrccm.154.6.8970383

Westra, J. W., Schlage, W. K., Frushour, B. P., Gebel, S., Catlett, N. L., Han, W., et al. (2011). Construction of a computable cell proliferation network focused on non-diseased lung cells. *BMC Syst. Biol.* 5:105. doi: 10.1186/1752-0509-5-105

Westra, J. W., Schlage, W. K., Hengstermann, A., Gebel, S., Mathis, C., Thomson, T., et al. (2013). A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue. *Bioinform. Biol. Insights* 7, 167–192. doi: 10.4137/BBI.S11509

Wheatley, D. N. (1995). Primary cilia in normal and pathological tissues. *Pathobiology* 63, 222–238. doi: 10.1159/000163955

Wills-Karp, M. (2004). Interleukin-13 in asthma pathogenesis. *Immunol. Rev.* 202, 175–190. doi: 10.1111/j.0105-2896.2004.00215.x

Workman, A. D., and Cohen, N. A. (2014). The effect of drugs and other compounds on the ciliary beat frequency of human respiratory epithelium. *Am. J. Rhinol. Allergy* 28, 454–464. doi: 10.2500/ajra.2014.28.4092

Wyatt, T. A., Spurzem, J. R., May, K., and Sisson, J. H. (1998). Regulation of ciliary beat frequency by both PKA and PKG in bovine airway epithelial cells. *Am. J. Physiol.* 275, L827–L835. doi: 10.1152/ajplung.1998.275.4.L827

Yaghi, A., and Dolovich, M. B. (2016). Airway Epithelial Cell Cilia and Obstructive Lung Disease. *Cells* 5:40. doi: 10.3390/cells5040040

Yaghi, A., Zaman, A., Cox, G., and Dolovich, M. B. (2012). Ciliary beating is depressed in nasal cilia from chronic obstructive pulmonary disease subjects. *Respir. Med.* 106, 1139–1147. doi: 10.1016/j.rmed.2012.04.001

Yang, B., Schlosser, R. J., and McCaffrey, T. V. (1996). Dual signal transduction mechanisms modulate ciliary beat frequency in upper airway epithelium. *Am. J. Physiol.* 270, L745–L751. doi: 10.1152/ajplung.1996.270.5.L745

Zagoory, O., Braiman, A., Gheber, L., and Priel, Z. (2001). Role of calcium and calmodulin in ciliary stimulation induced by acetylcholine. *Am. J. Physiol. Cell Physiol.* 280, C100–C109. doi: 10.1152/ajpcell.2001.280.1.C100

Zagoory, O., Braiman, A., and Priel, Z. (2002). The mechanism of ciliary stimulation by acetylcholine: roles of calcium, PKA, and PKG. *J. Gen. Physiol.* 119, 329–339. doi: 10.1085/jgp.20028519

Conflict of Interest Statement: All authors are employees of Philip Morris International.

Copyright © 2019 Yepiskoposyan, Talikka, Vavassori, Martin, Sewer, Gubian, Luettich, Peitsch and Hoeng. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 21 February 2019
doi: 10.3389/fgene.2019.00059

Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome

Ryan A. Miller[1†], Friederike Ehrhart[1,2†], Lars M. T. Eijssen[1,3], Denise N. Slenter[1], Leopold M. G. Curfs[2], Chris T. Evelo[1,2,4], Egon L. Willighagen[1] and Martina Kutmon[1,4]*

[1] Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands, [2] GKC-Rett Expertise Centre, MUMC+, Maastricht, Netherlands, [3] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands, [4] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands

OPEN ACCESS

Edited by:
Marco Antoniotti,
Università degli studi di Milano Bicocca, Italy

Reviewed by:
Zhi-Ping Liu,
Shandong University, China
Daniela Albrecht-Eckardt,
BioControl Jena GmbH, Germany

***Correspondence:**
Martina Kutmon
martina.kutmon@maastrichtuniversity.nl

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *11 November 2018*
Accepted: *24 January 2019*
Published: *21 February 2019*

Citation:
Miller RA, Ehrhart F, Eijssen LMT, Slenter DN, Curfs LMG, Evelo CT, Willighagen EL and Kutmon M (2019) Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. Front. Genet. 10:59. doi: 10.3389/fgene.2019.00059

Pathway and network approaches are valuable tools in analysis and interpretation of large complex omics data. Even in the field of rare diseases, like Rett syndrome, omics data are available, and the maximum use of such data requires sophisticated tools for comprehensive analysis and visualization of the results. Pathway analysis with differential gene expression data has proven to be extremely successful in identifying affected processes in disease conditions. In this type of analysis, pathways from different databases like WikiPathways and Reactome are used as separate, independent entities. Here, we show for the first time how these pathway models can be used and integrated into one large network using the WikiPathways RDF containing all human WikiPathways and Reactome pathways, to perform network analysis on transcriptomics data. This network was imported into the network analysis tool Cytoscape to perform active submodule analysis. Using a publicly available Rett syndrome gene expression dataset from frontal and temporal cortex, classical enrichment analysis, including pathway and Gene Ontology analysis, revealed mainly immune response, neuron specific and extracellular matrix processes. Our active module analysis provided a valuable extension of the analysis prominently showing the regulatory mechanism of *MECP2*, especially on DNA maintenance, cell cycle, transcription, and translation. In conclusion, using pathway models for classical enrichment and more advanced network analysis enables a more comprehensive analysis of gene expression data and provides novel results.

Keywords: pathway analysis, WikiPathways, Reactome, Rett syndrome, network analysis, RDF, topology, active subnetworks

1. INTRODUCTION

In a diseased state, many molecular processes in the human body are affected and dysregulated. Performing pathway analysis on molecular data sets comparing healthy vs. diseased subjects is immensely effective in finding affected pathways and it enables researchers to study the underlying processes in detail, to reveal possible disease mechanisms. While standard enrichment methods

have limitations and pathways are analyzed independently with their arbitrary process boundaries (Khatri et al., 2012), the pathway models themselves are very interesting from a network science perspective. These models contain detailed information about biological molecules and their interactions with one another, which can be visualized and analyzed using network biology tools (Kutmon et al., 2014). The detailed models of these biological processes are collected in online pathway databases like WikiPathways (Slenter et al., 2017) and Reactome (Fabregat et al., 2017). The availability of pathway models in the structured and semantic Resource Description Framework format (RDF) creates the possibility to integrate all pathway models into one large network and therefore incorporate the relations and overlap between them (Waagmeester et al., 2016). By removing artificial boundaries, this will enable us to study the systemic effects of diseases, such as Rett syndrome, using network biology methods. Specifically, we can look for subnetworks, even if not present in pathways as found in pathway databases, which reflect modules of differential biological activity.

Rett syndrome (MIM: 312750, Rett, 1966) is a rare genetic disorder, caused in most patients by a loss of function mutation in the *MECP2* gene (Amir et al., 1999). The accompanying MECP2 protein is multifunctional and acts as an epigenetic repressor, transcriptional repressor, and transcriptional activator. MECP2 binds DNA on methylated CpG islands and is involved in several regulatory activities: attracting histone deacetylases (HDAC1), increasing packing density of DNA, repressing and in specific genes also activating gene expression, and due to its phosphorylation sites, MECP2 activity is sensitive to intracellular signaling (Chunshu et al., 2006; Ehrhart et al., 2016). Due to its regulatory role, many downstream genes are affected in case of loss of function, resulting in a broad range of symptoms including moderate to severe intellectual disability, gait problems, stereotypic movements, dystonia, scoliosis, epileptic seizures, and sleep problems (Hagberg et al., 2002; Neul et al., 2010). In the past 10 years, omics data analysis on the level of genome, transcriptome, or proteome saw an increase in importance, to analyse and understand the holistic impact of MECP2, respectively, the impact of an impaired MECP2. Shovlin and Tropea (2018) recently reviewed the available transcriptomics studies on Rett syndrome and came to the conclusion that the most researched impact of MECP2 dysfunction lies with dendritic connectivity and synapse maturation, mitochondrial dysfunction, and glial cell activity. Recent pathway analysis results of single and integrated studies identified changes in intracellular signaling, including EIF2 (eukaryotic translation initiation) signaling, cytoskeleton, and cell metabolism including mitochondrial function (Bedogni et al., 2014; Ehrhart et al., 2018) .

In this study, we aim to investigate the molecular changes in Rett syndrome patients using a network-based approach by integrating existing pathway models from WikiPathways and Reactome into one large network and identifying disease-affected submodules that show differential gene expression. We will compare the results with standard enrichment analysis methods, including pathway and Gene Ontology analysis, and expect that the identified disease modules will also contain interactions in pathways not found through those methods.

2. MATERIALS AND METHODS

Dataset

The publicly available dataset studying the transcriptome in human brain tissue of Rett syndrome patients and healthy controls from the Gene Expression Omnibus (GEO) was used (GEO:GSE75303). The original study was published by Lin et al. (2016). The dataset contains transcriptome data obtained with Illumina HumanHT-12 V4.0 expression beadchips. The samples were taken postmortem from human frontal and temporal cortex of three Rett syndrome patients (*MECP2* mutations c.378-2A>G, c.763C>T, c.451G>T) and three age-, gender-, and ethnicity-matched controls.

Raw and normalized data as well as study metadata were obtained (GEO:GSE75303) and subjected to quality control, including signal distribution and sample grouping analyses, using plotting functions from ArrayAnalysis.org (Eijssen et al., 2013). No samples were excluded for further analysis. The provided normalized data on GEO was filtered to remove all probes with a detection p-value of 1 for all samples, indicating overall absence of expression. Thereafter, the limma package for R (version 3.30.13, Ritchie et al., 2015) was used to compute differential expression between Rett patients and controls for the frontal and temporal cortex samples separately. For each probe, this results in estimates of the fold change and p-value significance between the patient and control groups. Probes were re-annotated with Ensembl gene identifiers based on Ensembl build 91 using the BridgeDbR package (version 1.16.0, Leemans et al., 2018) with the Hs_Derby_Ensembl_91.bridge database (van Iersel et al., 2010).

Enrichment Analysis

We performed pathway analysis with PathVisio (version 3.3.0, Kutmon et al., 2015) and Gene Ontology (GO) analysis with GO-Elite (version 1.2, Zambon et al., 2012).

For GO analysis with GO-Elite, the input gene lists for frontal and temporal cortex contained all significantly changed genes (p-value < 0.05) with an absolute fold change cutoff of 1.5. Ensembl identifiers of all measured genes in the datasets were provided as the background list. Number of permutations was set to 2,000. Pruned GO-term results (i.e., GO terms for which genes in subterms that were found to be significant were removed) were filtered based on Z-score (> 1.96), permuted p-value (< 0.05) and a minimum number of changed genes of five. Pathway analysis was performed on a combined human pathway collection from all curated WikiPathways pathways including the Reactome pathway set (in total 903 pathways, October 2018 release). Differential gene expression was mapped to genes on the pathway diagrams using the Hs_Derby_Ensembl_91.bridge identifier mapping database. Thereafter, pathway statistics was performed on differential gene expression for temporal and frontal cortex using the following criteria to select only significantly differentially expressed genes (absolute fold change cutoff of 1.5 and p-value < 0.05):

`(log2FC < -0.58 OR log2FC > 0.58) AND p-value < 0.05.`

The resulting ranked pathway list was filtered based on Z-score (> 1.96), permuted p-value (< 0.05), and minimum number of changes (positive) genes of five.

Pathway-Based Network Construction

Biological pathway models are small sub-networks describing specific biological processes. Connecting and integrating pathway models in one large network enables us to use network biology tools and approaches to study and investigate the network.

We used the WikiPathways RDF from October 2018 release (Waagmeester et al., 2016) to retrieve information about all interactions in the pathway models of two major pathway databases, WikiPathways and Reactome. With this network approach, the pathway models are not treated as independent modules, but they are integrated on an interaction level, which enables linking information from different pathways based on their shared participants and thus bringing related interactions closer to each other. As shown in **Figure 1**, each interaction is represented by an interaction node in the network with edges to all participant nodes (either source, target, or participant). For each interaction, it is recorded in which pathway or pathways the interaction is present. By connecting all the retrieved interactions, a large network representing all human pathway models was created. The SPARQL query language was used to retrieve the relevant data. The scripts to generate the constructed network are available on GitHub (https://github.com/wikipathways/wprdf2cytoscape). Interactions with at least two annotated interaction participants (gene product, metabolite, complex) are included. Gene products have unified Ensembl (Zerbino et al., 2017) identifiers, metabolites have either Wikidata (Mietchen et al., 2015), ChEBI (Hastings et al., 2015) or HMDB identifiers (Wishart et al., 2017), and complexes have Reactome identifiers. A list of frequently occurring small molecules (**Supplementary Table 1**), e.g., H^+, H_20, ATP, were removed from the network to prevent inclusion of paths with no specific biological relevance. Such small molecules tend to create artificial hub nodes simply because e.g., ATP is used/produced in a lot of metabolic reactions.

Active Module Analysis

The constructed network was loaded into Cytoscape (version 3.7.0), a network analysis and visualization tool (Shannon et al., 2003). Differential expression analysis data (log2 fold changes and p-values) for both frontal and temporal cortex were added as node attributes to the network.

The Cytoscape app jActiveModules (version 3.2.1, Ideker et al., 2002) was used to identify active submodules in the large network that show significant changes in expression. These subnetworks are freed from the artificial pathway boundaries of conventional pathway models found in WikiPathways and Reactome. The following parameters were used to find active submodules: p-value as the node attribute, number of modules was set to five, overlap threshold of 0.8, and search strategy with a search depth of two.

Tools and Settings

- Dataset: Normalized data from GEO, plotting functions from ArrayAnalysis.org, limma package for R (version 3.30.13), BridgeDbR package (version 1.16.0) with Hs_Derby_Ensembl_91.bridge database.

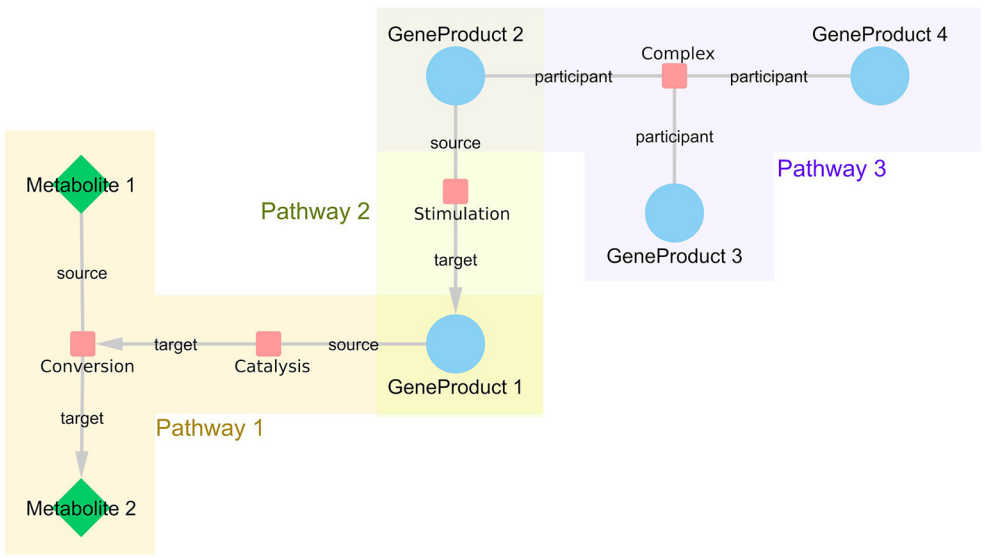


FIGURE 1 | WikiPathways network structure. Every interaction is represented as a node in the network with links to all participants. If the interaction is directed, the information about source and target nodes is added as an edge attribute. The nodes represented as small, red rounded rectangles are interactions, blue circles represent gene products and green diamonds embody metabolites. Interactions that share certain participants, such as GeneProduct 1, are brought close together in the resulting network even if they are from different pathways, such as Pathway 1 and 3.

- Enrichment analysis: PathVisio (version 3.3.0) and GO-Elite (version 1.2).
- Pathway-based network construction: Script available on Github (https://github.com/wikipathways/wprdf2cytoscape)
- Active module analysis: Cytoscape (version 3.7.0), jActiveModules app (version 3.2.1)

3. RESULTS

Gene Expression

The total number of probes measured was 37,707 from which 29,024 could be linked to Ensembl identifiers. After merging multiple probe identifiers for the same Ensembl identifier, 19,023 unique gene identifiers remained. Differential gene expression analysis revealed 1,953 in the frontal cortex and 2,436 significantly changed genes in the temporal cortex samples of RETT syndrome patients vs. controls. Only 221 in frontal and 341 of the significantly changed genes in temporal cortex had a more than 1.5-fold increase or decrease in expression (|log2 fold change| $>$ 0.58). In both brain regions, more genes were down-regulated in Rett syndrome patients than up-regulated, see **Table 1**, which matches with findings from the original publication (Lin et al., 2016).

Gene Ontology Analysis

Gene Ontology overrepresentation analysis identified 39 and 50 biological processes as altered in frontal and temporal cortex, respectively (**Supplementary Tables 2**, **3**). Summarizing, neuron specific and immune system-related processes were found to be enriched in both brain regions for Rett syndrome patients. In temporal cortex, additionally, regulation of translational initiation (GO:0006446) and an extracellular matrix/cytoskeleton-related process (GO:0007229) were found to be enriched. Interestingly, the microglia relevant complement factors C1QB and C1QC were found in the enriched GO classes defense response (GO:0006952) and immune effector process (GO:0002252).

Pathway Analysis

Pathway analysis was performed in PathVisio for both brain regions separately. Overrepresentation analysis revealed 18 and 21 pathways altered in the datasets for frontal and temporal cortex, respectively (Z-score $>$ 1.96, minimum five changed genes), see **Figure 2**. Interestingly, eight pathways were altered in both frontal and temporal cortex. Similar to the results of the GO analysis, several immune system-related and extracellular matrix/cytoskeleton-related pathways were found to be enriched. Additionally, calcium channel related processes including muscle contraction pathways were found in both brain regions. Although muscle contraction pathways are not expected in brain tissue samples, the overlapping differentially expressed genes were mostly ion channels and signaling cascade proteins also highly relevant for neurons. **Supplementary Figure 1** shows the heatmap with a more lenient filter (Z-score $>$ 1.96, minimum three changed genes). **Figure 3** is an example pathway visualization for a pathway that has a high Z-score in both tissue types, Microglia Pathogen Phagocytosis Pathway (Hanspers and Slenter, 2017).

Pathway-Based Network Construction

From the 903 pathway models in the WikiPathways and Reactome collection, 860 pathways contained 27,410 unique interactions. On average, a pathway contained 35 interactions (min = 1, max = 510, median = 22). Interestingly, 3,264 interactions occur multiple times but only 2,103 interactions are present in more than one pathway. As an example, one of the highest occurring interactions is the complex binding of the three subunits of the IκB kinase complex which plays an important role in the propagation of cellular response to inflammation (Häcker and Karin, 2006) and is present in 25 different pathways.

The resulting network consists of 48,639 nodes and 106,137 edges. The network consists of one major component (46,756 nodes) and 427 smaller components with each less than twenty nodes. The network contains 8,643 gene products, 2,704 metabolites and 9,882 complex / group nodes. Most common interaction types are directed interaction (13,572), complex / group participation (5,298), catalysis (4,787), inhibition (1,185), and conversions (896).

TABLE 1 | Differentially expressed genes in frontal and temporal cortex.

	Temporal cortex down-regulated	Temporal cortex not changed	Temporal cortex up-regulated
Frontal cortex down-regulated	88 ↓↓	44 ↓-	1 ↓↑
Frontal cortex not changed	171 -↓	18,576 - -	55 -↑
Frontal cortex up-regulated	3 ↑↓	62 ↑-	23 ↑↑

133 and 88 genes were significantly down- and up-regulated in frontal cortex, respectively. Two hundred sixty-two and 79 genes were significantly down- and up-regulated in temporal cortex, respectively. Eighty-eight genes are down-regulated, and 23 genes are up-regulated in both brain regions. Only four genes show different expression patterns. The following filtering criteria were used: p-value $<$ 0.05 and absolute log2 fold change $>$ 0.58.

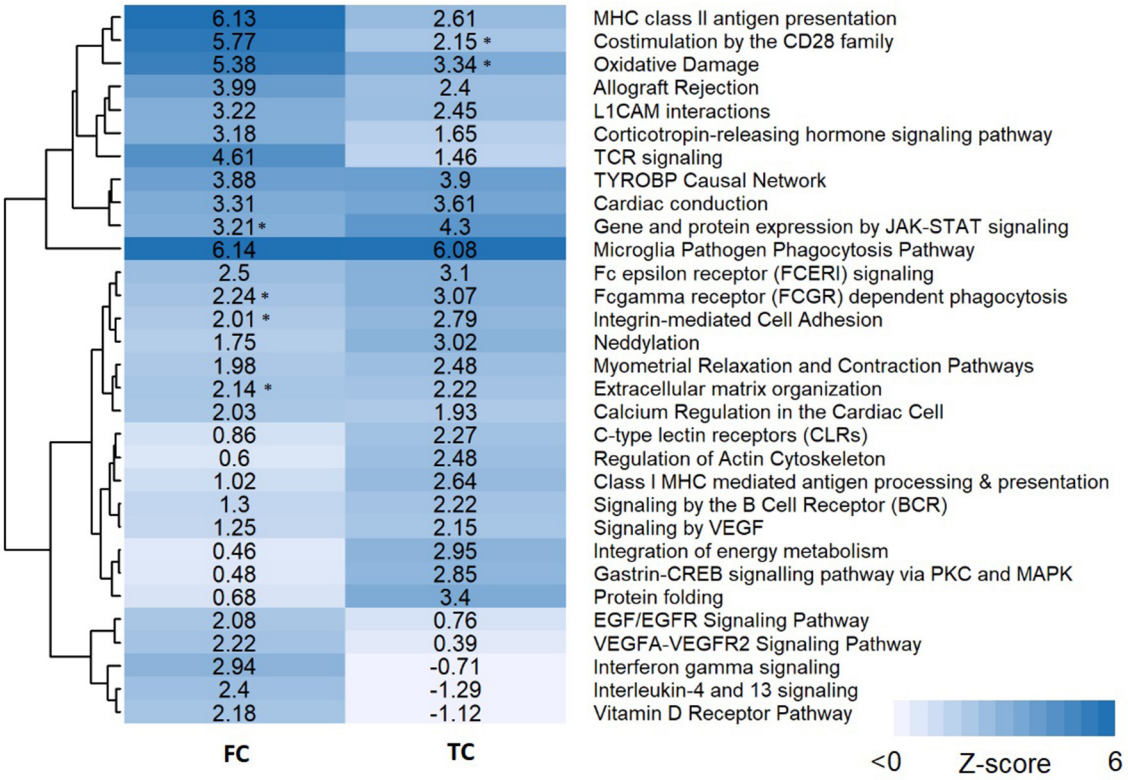


FIGURE 2 | Pathway analysis results for frontal and temporal cortex data. Pathways are clustered in this heatmap based on their Z-scores. Pathways with a high Z-score (>1.96) contain significantly more changed genes than expected and are considered pathways of interest. An asterisk next to the Z-score value indicates pathways with a significant Z-score (>1.96) but less than five changed genes.

Active Module Analysis

Active modules were calculated using the jActiveModules app. The top five modules with the highest active paths scores were identified for both comparisons, frontal and temporal cortex. The modules for frontal cortex contained between 300–350 nodes and 560–1,020 edges. The top modules for temporal cortex tended to be smaller ranging from 230–290 nodes and 450–1,000 edges. **Figures 4**, **5** show the highest-ranked module for frontal and temporal cortex, respectively. Gene expression changes are visualized as node color and significance is indicated by the node border color. All modules only contained gene products; no metabolites were found. The complete submodule analysis results for both datasets can be found in **Supplementary Data 1** (zip file containing two Cytoscape session files).

The highest ranked active module for frontal cortex contains 303 nodes (79 interactions and 224 gene products) and 568 edges, see **Figure 4**. Two hundred and ten of the gene products are measured in the dataset and 112 are changed significantly (p-value < 0.05). Twelve gene products have an absolute log2 fold change > 0.58. The subnetwork contains eight significantly down-regulated genes (blue rounded rectangles) including two F-Box genes, *FBOX32* and *FBXO9*, involved in phosphorylation-dependent ubiquitination. The subnetwork contains five significantly up-regulated genes (red rounded rectangles) with diverse roles. The genes identified as hubs in the active module network of frontal cortex are two gene products which are not measured in the dataset, *RPS27A* and *UBA52*. Both are involved in protein degradation via 26S proteasome, ubiquitination, translation, and DNA excision repair. In the central part of the network, the ribosomal proteins including *RPL14*, *RPL29*, and *RPL3* form a cluster. This cluster is connected via *PPP2CA* and *PPP2R1A*, two phosphatases involved in cell cycle, DNA replication and transcription, to a cluster of centrosomal proteins including *CEP78*, *CEP57*, and *CEP131*. The module combines interactions from 47 unique pathways (**Supplementary Table 4**) including class I MHC mediated antigen processing and presentation (WP3577), non-sense-mediated decay (WP2710), cell-cycle related pathways (WP1859, WP1775, WP1858, WP2772), and eukaryotic translation elongation and initiation (WP1811, WP1812).

The highest ranked active module for temporal cortex contains 238 nodes (84 interactions and 154 gene products) and 457 edges, see **Figure 5**. The module partially overlaps with the module found for frontal cortex. One hundred and fourty three of the gene products are measured in the dataset and 137 are changed significantly (p-value $<$ 0.05). Twenty-nine gene products have an absolute log2 fold change > 0.58. The module contains 24 significantly

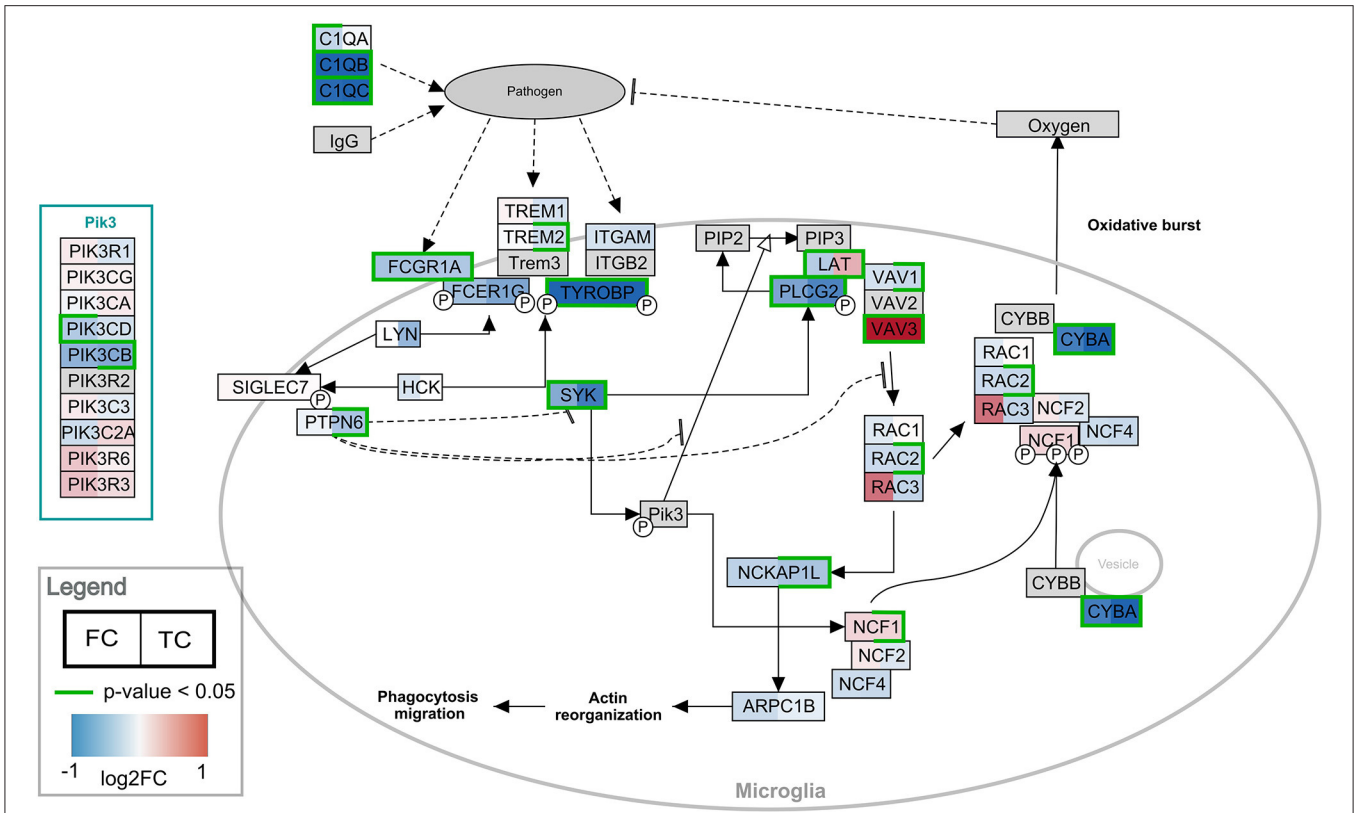


FIGURE 3 | Visualization of the frontal and temporal cortex gene expression on the Microglia Pathway Phagocytosis Pathway. In the left half of the gene boxes, the gene expression change in the frontal cortex is shown. In the right half of the gene boxes, the gene expression in the temporal cortex is shown. The blue colors represent down-regulation of the gene in Rett syndrome patients (negative log2 fold change), while the red shades visualize the up-regulated genes. The darker the color, the stronger the effect. Green borders indicate significance of the change (p-value < 0.05). Gray colored nodes are not annotated or measured in the dataset.

down-regulated genes (blue rounded rectangles) including several ubiquitin conjugating enzymes (*UBE2E1*, *UBE2E3*) and translation initiation factors (*EIF4A2*, *EIF4H*, *EIF4G2*). Only five significantly up-regulated genes are found in the subnetwork (red rounded rectangles) but the distance between them is large. This subnetwork contains similar hub nodes as in the frontal cortex subnetwork including *RPS27A*, *UBA52*, and *PPP2R1A*. Additionally, *NCBP2* and *NCBP1*, proteins involved in RNA processing, play an important role in the subnetwork. The module combines interactions from 51 unique pathways (**Supplementary Table 5**) including transcription / translation (WP1889, WP1906, WP1812), cell cycle (WP1859, WP1775, WP4109), and immune response (WP3577, WP2658) related processes.

4. DISCUSSION

MECP2 is a multifunctional protein which is involved in several transcriptional inhibitory and activational processes. MECP2 was generally regarded as a repressor, however its role as genetic activator has also been confirmed (Chahrour et al., 2008). In previous studies, a loss of function in MECP2 due to a mutation has been found to influence a variety of pathways and biological processes, including pathways related to not only neuron development and function, but also to the immune system, transcription, and translation related processes (which were identified mainly by transcriptome analysis, Colantuoni et al., 2001; Bedogni et al., 2014; Ehrhart et al., 2018; Shovlin and Tropea, 2018). The affected pathways identified with our study closely match the results previously found by Ehrhart et al. (2018), in which human brain tissue data of Rett syndrome patients (published by Deng et al., 2007) was analyzed. The expression of the MECP2 protein itself is not significantly affected in this dataset (minor, insignificant down regulation, log2 fold change of –0.1, in both brain regions).

The original study by Lin et al. (2016) from which the dataset analyzed in this paper was acquired, focused on the significant down-regulation of certain complement system factors in Rett syndrome (C1QA, C1QB, C1QC). Complement system factors are produced generally in liver, however their expression was also found to be changed in stimulated microglia. Furthermore, there is emerging evidence that C1Q factors are involved in several non-immunogenic activities, such as synaptic pruning in microglia (Kouser et al., 2015).

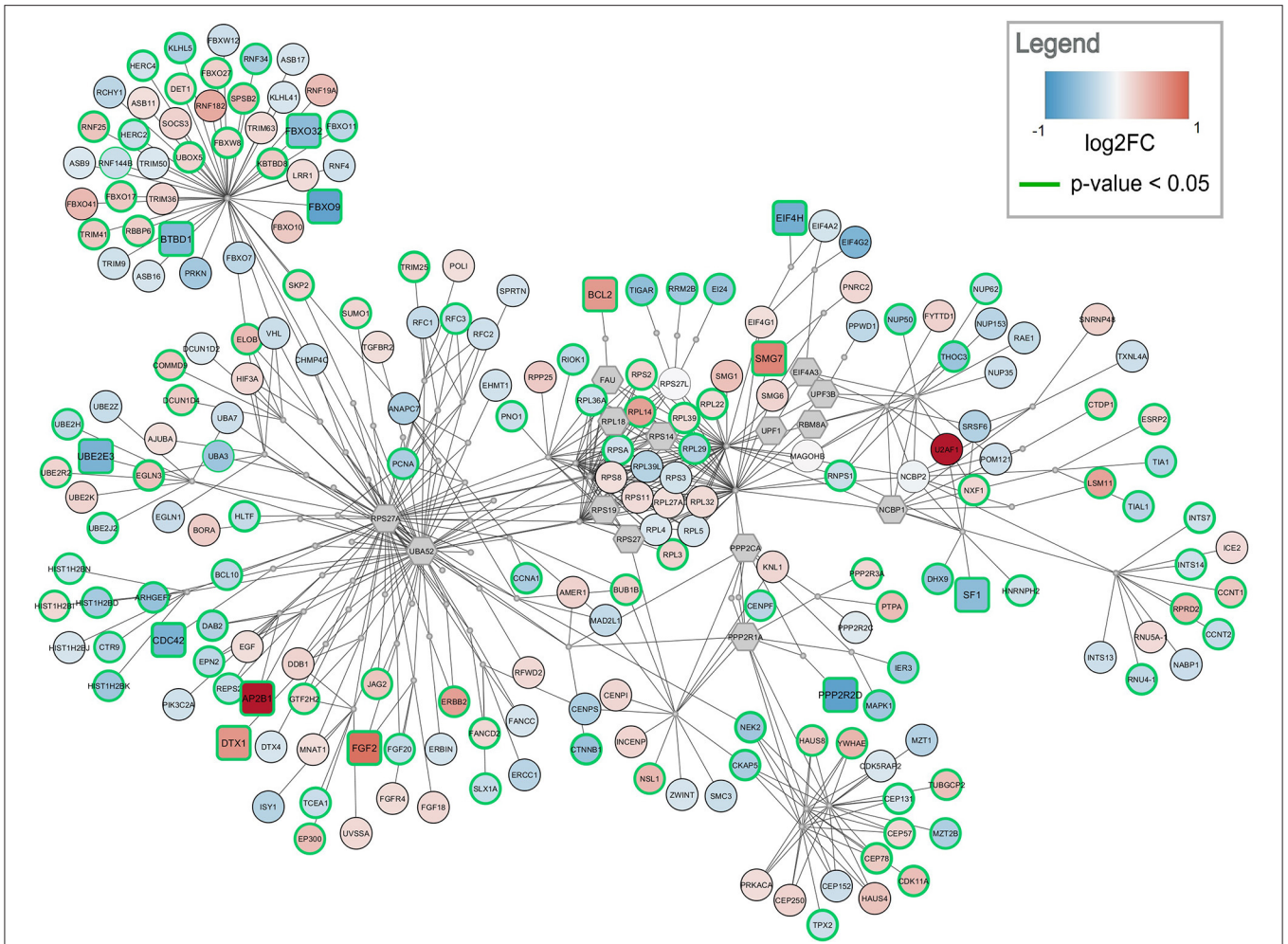


FIGURE 4 | Top-ranked active module for frontal cortex data. The highest-ranked subnetwork contains 303 nodes and 568 edges. It contains 13 significantly changed genes (rounded rectangles) when applying the same cutoff as for enrichment analysis (absolute log2 fold change > 0.58). Other measured gene products are visualized as circular nodes. Blue fill color indicates down-regulation while red indicates up-regulation. The darker the color, the stronger the effect. Gray hexagons are gene products not measured in the data set. The very small, gray nodes represent interaction nodes. These were combined from 47 different pathways, with none of the pathways providing more than six interactions.

As expected, our pathway and GO analysis revealed a substantial number of immune system related pathways to be affected in Rett syndrome frontal and temporal cortex tissue samples. Inflammatory processes have been identified previously in Rett syndrome patients, mouse models and *in vitro* systems, and are suspected to contribute to the development of Rett syndrome (De Felice et al., 2016; Ehrhart et al., 2018). **Figure 2** shows many of affected pathways in both frontal and temporal cortex, with similar results found by GO analysis. Interestingly, no complement system or transcription / translation related pathways show up (except Microglia Pathogen Phagocytosis Pathway, which includes C1Q factors). Only seven of the 31 pathways found through pathway analysis contribute interactions to the active modules identified for frontal and temporal cortex. The modules mainly contained interactions from transcription / translation and cell cycle related pathways, which were not found with the classical enrichment analysis. These processes were also previously found in transcriptome pathway analysis of Rett syndrome (Bedogni et al., 2014; Ehrhart et al., 2018). Not surprisingly, the subnetworks do not contain metabolic reactions. Only metabolites connecting at least two genes affected by MECP2 would be present in an active subnetwork. The enrichment analysis did not show any metabolic processes that are affected, which is in line with the manifestation of Rett syndrome. Overall, the regulatory effects of MECP2, especially on DNA maintenance, cell cycle, transcription, and translation, is more prominently shown in the active modules, while immune system related responses are more present in pathway analysis. Importantly, the active module approach does not replace analyses like classical enrichment analysis but augments it. When running the active module analysis on the same

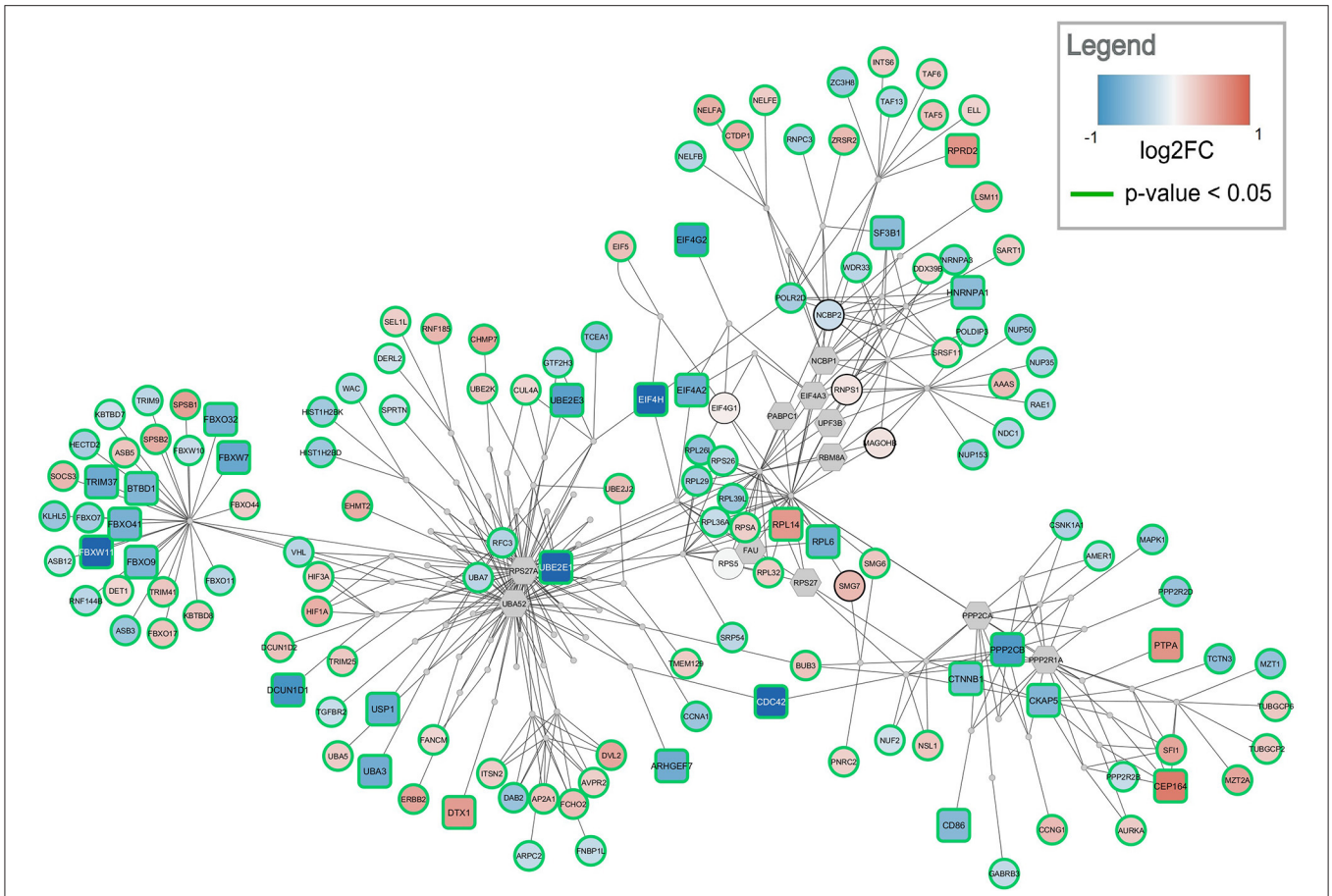


FIGURE 5 | Top-ranked active module for temporal cortex data. The subnetwork contains 238 nodes and 457 edges. It contains 29 significantly changed genes (rounded rectangles) when applying the same cutoff as for enrichment analysis (absolute log2 fold change > 0.58). Other measured gene products are visualized as circular nodes. Blue fill color indicates down-regulation while red indicates up-regulation. The darker the color, the stronger the effect. Gray hexagons are gene products not measured in the data set. The very small, gray nodes represent interaction nodes. These were combined from 51 different pathways, with none of the pathways providing more than six interactions.

network using the dataset with permuted gene labels, the resulting subnetworks are very different from the identified Rett subnetworks. This basic computational validation further strengthens our confidence that we indeed have subnetworks specific and strongly affected in Rett syndrome patients. The results of the permutation analysis are summarized in **Supplementary Data 2.**

This was the first time the entirety of the WikiPathways knowledgebase, including Reactome pathways, has been used to create a comprehensive human pathway-based network for network analysis of transcriptomics data. While the pathway content of both databases overlaps, both resources also contain unique information. By building a network out of pathways from a combination of pathway databases, a more complete biological (and therefore genome) coverage is enabled. Identifying active modules from a large network has some major benefits, such as the easy applicability to any gene expression dataset, ignoring predefined boundaries used in traditional pathway diagrams, and incorporating the relations and overlap between the pathways. Additionally, this method does not require researchers to predefine a certain cutoff, since genes are ranked based on their significance.

Some considerations arose when constructing and analyzing the network. For instance, some common metabolites like ATP, ADP, or NADH, while biologically necessary, were excluded from the network, since their involvement in a multitude of interactions created links between almost every node. Additionally, this approach is strongly depending on the a priori input of pathway data in terms of coverage and quality. Currently, human pathway databases contain a little over 50% of the protein coding genes (Slenter et al., 2017), which is also a probable number for the coverage of metabolites and interactions. Pathway models generally contain information about directionality of the interactions. However, available active subnetwork analysis methods only take topology but not directionality into account. This could strongly affect the identification of active submodules and would be an important extension of existing algorithms.

The active module discovery approach should be considered as an additional step after classical enrichment analysis. In this

study, we used human brain transcriptomics data from a study with Rett syndrome patients, however our approach is not unique to this application or rare diseases. These diseases are by definition less common and often less extensively studied, which may result in lower availability of specific pathway models. Nonetheless, the active module approach succeeds and shows great power for additional discoveries. While rare genetic diseases have the advantage that the causative gene is (usually) known, the resulting downstream consequences can be diverse and affect a variety of pathways. By using pathway models in an integrative network approach, further use of the invaluable resources present in pathway databases is enabled and subnetworks of interest can be retrieved based on the entire body of pathways available. Using Cytoscape allows using all available apps such as the jActiveModules app to analyse our network. Importantly, the complete interaction network of WikiPathways with 48,639 nodes and 106,137 edges can be opened and analyzed in Cytoscape, despite of the network to be too large to be visualized. The use of graph databases like Neo4j, which already have connections available to Cytoscape (cyNeo4j app, Summer et al., 2015), could be a useful addition to the approach. Importantly, as part of the systems biology cycle, advanced computational analyses like the one reported in this manuscript lead to new hypotheses and ideas for experiments, which then need to be tested and validated in a laboratory.

Conclusion

Pathway models have proven themselves as powerful tools for biologists to describe and analyse biological processes. The collaboration between the widely-adopted pathway databases WikiPathways and Reactome and the availability of their data in RDF format allowed us to integrate a large number of pathways from both databases into one large network. This enables us to perform advanced network analyses like active submodule identification. By comparing classical enrichment methods with the active submodule identification on a Rett syndrome dataset in two different brain regions, we found that both approaches provided valuable insights into the disease. Importantly, they were strongly complementary and did not show the same results.

DATA AVAILABILITY STATEMENT

The dataset analyzed for this study can be found in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75303).

AUTHOR CONTRIBUTIONS

RM and FE: data analysis, literature search; LE: data preprocessing and statistical analysis; DS: data analysis; LC: literature search; CE and EW: study design; MK: data analysis, study design, literature search. All authors contributed in writing and editing of the manuscript.

FUNDING

FE, DS, CE, and EW received funding from ELIXIR, the European research infrastructure for life-science data for development of interoperability approaches used in this study. FE is funded by Stichting Terre, the Dutch Rett syndrome fonds. This project was partly funded by the Dutch Province of Limburg (MK). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 825575, EJP-RD (FE, CE).

ACKNOWLEDGMENTS

The authors would like to thank the authors and curators of WikiPathways and Reactome for the provision of content. Special thanks to Lin et al., the authors of the original publication, for generating the data and making it available for re-use. Thanks to Irina Voineagu for responding to our inquiries and answering our questions regarding the dataset.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00059/full#supplementary-material

REFERENCES

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* 23:185. doi: 10.1038/13810

Bedogni, F., Rossi, R. L., Galli, F., Cobolli Gigli, C., Gandaglia, A., Kilstrup-Nielsen, C., et al. (2014). Rett syndrome and the urge of novel approaches to study MeCP2 functions and mechanisms of action. *Neurosci. Biobehav. Rev.* 46, 187–201. doi: 10.1016/j.neubiorev.2014.01.011

Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T., Qin, J., et al. (2008). MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* 320, 1224–1229. doi: 10.1126/science.1153252

Chunshu, Y., Endoh, K., Soutome, M., Kawamura, R., and Kubota, T. (2006). A patient with classic Rett syndrome with a novel mutation in MECP2 exon 1. *Clin. Genet.* 70, 530–531. doi: 10.1111/j.1399-0004.2006.00712.x

Colantuoni, C., Jeon, O.-H., Hyder, K., Chenchik, A., Khimani, A. H., Narayanan, V., et al. (2001). Gene expression profiling in postmortem Rett syndrome brain: differential gene expression and patient classification. *Neurobiol. Dis.* 8, 847–865. doi: 10.1006/nbdi.2001.0428

De Felice, C., Leoncini, S., Signorini, C., Cortelazzo, A., Rovero, P., Durand, T., et al. (2016). Rett syndrome: an autoimmune disease? *Autoimmun. Rev.* 15, 411–416. doi: 10.1016/j.autrev.2016.01.011

Deng, V., Matagne, V., Banine, F., Frerking, M., Ohliger, P., Budden, S., et al.. (2007). FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice. *Hum. Mol. Genet.* 16, 640–650. doi: 10.1093/hmg/ddm007

Ehrhart, F., Coort, S. L., Cirillo, E., Smeets, E., Evelo, C. T., and Curfs, L. M. (2016). Rett syndrome–biological pathways leading from MECP2 to disorder phenotypes. *Orphanet J. Rare Dis.* 11, 158. doi: 10.1186/s13023-016-0545-5

Ehrhart, F., Coort, S. L., Eijssen, L., Cirillo, E., Smeets, E., Sangani, N. B., et al. (2018). Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. *bioRxiv*, 274258. doi: 10.1101/274258

Eijssen, L. M., Jaillard, M., Adriaens, M. E., Gaj, S., de Groot, P. J., Müller, M., et al. (2013). User-friendly solutions for microarray quality control

and pre-processing on ArrayAnalysis.org. *Nucleic Acids Res.* 41, W71–W76. doi: 10.1093/nar/gkt293

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2017). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Häcker, H., and Karin, M. (2006). Regulation and function of IKK and IKK-related kinases. *Sci. Stke* 2006:re13. doi: 10.1126/stke.3572006re13

Hagberg, B., Hanefeld, F., Percy, A., and Skjeldal, O. (2002). An update on clinically applicable diagnostic criteria in Rett syndrome. Comments to Rett syndrome clinical criteria consensus panel satellite to european paediatric neurology society meeting, Baden Baden, Germany, 11 September 2001. *Eur. J. Paediatr. Neurol.* 6, 293. doi: 10.1053/ejpn.2002.0612

Hanspers, K. and Slenter, D. (2017). *Microglia Pathogen Phagocytosis Pathway (Homo sapiens)*. WikiPathways. Available online at: https://www.wikipathways.org/instance/WP3937

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., et al. (2015). ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 44, D1214–D1219. doi: 10.1093/nar/gkv1031

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl_1), S233–S240. doi: 10.1093/bioinformatics/18.suppl_1.S233

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi: 10.1371/journal.pcbi.1002375

Kouser, L., Madhukaran, S. P., Shastri, A., Saraon, A., Ferluga, J., Al-Mozaini, M., et al. (2015). Emerging and novel functions of complement protein C1q. *Front. Immunol.* 6:317. doi: 10.3389/fimmu.2015.00317

Kutmon, M., Lotia, S., Evelo, C. T., and Pico, A. R. (2014). WikiPathways app for Cytoscape: making biological pathways amenable to network analysis and visualization. *F1000Res.* 3:152. doi: 10.12688/f1000research.4254.2

Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., et al. (2015). PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.* 11:e1004085. doi: 10.1371/journal.pcbi.1004085

Leemans, C., Willighagen, E., Bohler, A., and Eijssen, L. (2018). *BridgeDbR: Code for Using BridgeDb Identifier Mapping Framework From Within R.* R package, 1.16.0. doi: 10.18129/B9.bioc.BridgeDbR

Lin, P., Nicholls, L., Assareh, H., Fang, Z., Amos, T. G., Edwards, R. J., et al. (2016). Transcriptome analysis of human brain tissue identifies reduced expression of complement complex C1Q genes in Rett syndrome. *BMC Genomics* 17:427. doi: 10.1186/s12864-016-2746-7

Mietchen, D., Hagedorn, G., Willighagen, E., Rico, M., Gómez-Pérez, A., Aibar, E., et al. (2015). Enabling open science: wikidata for research (Wiki4R). *Res. Ideas Outcomes* 1:e7573. doi: 10.3897/rio.1.e7573

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, N., L., et al. (2010). Rett syndrome: revised diagnostic criteria and nomenclature. *Ann. Neurol.* 68, 944–950. doi: 10.1002/ana.22124

Rett, A. (1966). On a unusual brain atrophy syndrome in hyperammonemia in childhood. *Wiener Med Wochenschrift (1946)* 116, 723.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shovlin, S., and Tropea, D. (2018). Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet J. Rare Dis.* 13, 113. doi: 10.1186/s13023-018-0857-8

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46, D661–D667. doi: 10.1093/nar/gkx1064

Summer, G., Kelder, T., Ono, K., Radonjic, M., Heymans, S., and Demchak, B. (2015). cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics* 31, 3868–3869. doi: 10.1093/bioinformatics/btv460

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., et al. (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11:5. doi: 10.1186/1471-2105-11-5

Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E. L., Evelo, C. T., et al. (2016). Using the semantic web for rapid integration of WikiPathways with other biological online data resources. *PLoS Comput. Biol.* 12:e1004989. doi: 10.1371/journal.pcbi.1004989

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., et al. (2017). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 46, D608–D617. doi: 10.1093/nar/gkx1089

Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., et al. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–2210. doi: 10.1093/bioinformatics/bts366

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2017). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098

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.

Copyright © 2019 Miller, Ehrhart, Eijssen, Slenter, Curfs, Evelo, Willighagen and Kutmon. 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) and the copyright owner(s) 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 Genetics

REVIEW
published: 05 March 2019
doi: 10.3389/fgene.2019.00155

A Comprehensive Survey of Tools and Software for Active Subnetwork Identification

Hung Nguyen[1], *Sangam Shrestha*[1], *Duc Tran*[1], *Adib Shafi*[2], *Sorin Draghici*[2,3]* and *Tin Nguyen*[1]*

[1] Department of Computer Science and Engineering, University of Nevada, Reno, NV, United States, [2] Department of Computer Science, Wayne State University, Detroit, MI, United States, [3] Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, United States

A recent focus of computational biology has been to integrate the complementary information available in molecular profiles as well as in multiple network databases in order to identify connected regions that show significant changes under different conditions. This allows for capturing dynamic and condition-specific mechanisms of the underlying phenomena and disease stages. Here we review 22 such integrative approaches for active module identification published over the last decade. This article only focuses on tools that are currently available for use and are well-maintained. We compare these methods focusing on their primary features, integrative abilities, network structures, mathematical models, and implementations. We also provide real-world scenarios in which these methods have been successfully applied, as well as highlight outstanding challenges in the field that remain to be addressed. The main objective of this review is to help potential users and researchers to choose the best method that is suitable for their data and analysis purpose.

Keywords: active module, active subnetwork, subnetwork identification, data integration, PPI network, network analysis

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council (CNR), Italy

Reviewed by:
Alexey Sergushichev,
ITMO University, Russia
Rosalba Giugno,
University of Verona, Italy

***Correspondence:**
Sorin Draghici
sorin@wayne.edu
Tin Nguyen
tinn@unr.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *25 September 2018*
Accepted: *13 February 2019*
Published: *05 March 2019*

Citation:
Nguyen H, Shrestha S, Tran D, Shafi A, Draghici S and Nguyen T (2019) A Comprehensive Survey of Tools and Software for Active Subnetwork Identification. Front. Genet. 10:155. doi: 10.3389/fgene.2019.00155

1. INTRODUCTION

From human society to cellular activity, collaborative interactions, i.e., small units working in concert to accomplish certain functions, are an essential part of life. In complex multicellular organisms, their survival and health depend on the integrated activity of billions or trillions of cells organized into organ systems. Even in a single cell, the smallest structural and biological unit of life, fundamental processes, from DNA replication and energy production, to intercellular and intracellular signaling, often involve multiple biochemical reactions and molecular interactions taking place at multiple levels (transcriptomics, epigenomics etc.).

In order to have a good understanding of cellular functions at the systems-level, one needs to correctly identify and interpret all functional interactions of DNA, RNA, and proteins of organisms of interest (Szklarczyk et al., 2010). In turn, this has lead to the development of knowledge bases of functional modules and large networks of intermolecular interactions and pathways. Biological networks, which are graphical representation of genes, proteins, DNAs, RNAs, or even small miRNAs and their functional interactions, are rapidly accumulated in public databases, including HPRD (Keshava Prasad et al., 2008), DIP (Salwinski et al., 2004), KEGG (Kanehisa et al., 2017), Reactome (Croft et al., 2014), and many other curated interactome networks developed for human and model species (Harbison et al., 2004; Stelzl et al., 2005; Yu et al., 2008; Ravasi et al., 2010).

Many computational approaches have been developed to mine such interactome networks in order to better understand cellular processes and disease mechanisms (Spirin and Mirny, 2003). Topological modules (Girvan and Newman, 2002), within which nodes are well-connected and the interactions are more concentrated compared with those outside, are among the most intensively studied research areas. However, as functional interactions are annotated in static experimental conditions, network databases alone fail to account for the dynamic nature of biological systems and thus fail to provide a full representation of cellular interactions.

Recently, with the advancement of high-throughput technologies, biological data of different kinds have rapidly accumulated in public repositories. Taken alone, molecular data only represents a snapshot of biological systems and often fail to elucidate biological mechanisms. When projected onto biological networks, however, molecular profiles and expression changes have the potential to reflect the perturbation of complex cellular network and thus allow for comprehensive monitoring of biological systems (Cowen et al., 2017; Yi et al., 2017). A recent focus of computational biology has been to integrate the complementary information available in molecular profiles as well as in multiple network databases in order to identify active modules, i.e., well-connected subnetworks that are significantly perturbed under different conditions (Mitra et al., 2013). These approaches have been widely applied and proven to be powerful in elucidating biological mechanisms of underlying physiological and disease phenotypes (Chuang et al., 2007; Bapat et al., 2010; Qiu et al., 2010; Zhang and Ouellette, 2011; Shafi et al., 2019).

In this document, we categorize and review 22 such subnetwork identification methods based on the following criteria: their availability and user interface, the type of input the method requires, subnetwork seeding and construction, and statistical approaches used to assess the significance of the identified subnetworks. We classify these approaches into six different categories according to the techniques used to traverse the global network in order to construct the active subnetworks. In section 2, we discuss the availability, implementation, types of experimental input and reference network databases that the surveyed methods use. In section 3, we categorize and compare the methods according to the way they traverse and expand the subnetwork. In section 4, we include real-world scenarios in which the surveyed methods were successfully applied. In section 5, we discuss the limitations of current knowledge bases and outstanding challenges in method development. In section 6, we systematically recapitulate the 22 approaches by highlighting their key characteristics and differences. We also provide detailed descriptions for individual methods in **Supplementary Material**.

To the best of our knowledge, this is the first article that provides such in-depth discussion and covers a large number of tools for active subnetwork identification. A recent survey of biological networks (Mitra et al., 2013) discussed active network identification, among other topics, and provided a list of tools. However, this article covers many topics and its wide breadth means there was some limitation in the depth to which these tools could be covered. In addition, many of the tools listed there are outdated and/or not maintained anymore. More recently, another survey (He et al., 2017) focused on assessing the performance of 10 subnetwork analysis methods using simulations. This survey, however, provides even fewer details and discussion of each individual method. In contrast, here we provide a comprehensive review of a total of 22 methods for active subnetwork identification, highlighting their availability, implementation, applicable network databases, underlying mathematical and algorithmic principles, as well as advantages and limitations for each method. The main objective of this review is to help potential users and researchers to choose methods that are suitable for their data and analysis purpose.

2. SOFTWARE AND DATABASES

2.1. Availability and Implementation

Table 1 shows the 22 methods we review in this article. Although more computational methods for subnetwork identification have been published, we only review methods that are associated with executable packages that can actually be used by people other than the authors. This table provides the following information about each tool: (i) their availability (link to the tool), (ii) implementation (standalone package, web interface, user interface, programming language), (iii) reference to the original articles, (iv) citations, and (v) software license. We believe that these details are crucial for users to know before spending a significant amount of time to understand the software and perform analyses.

One often thinks that the strengths of a computational approach mostly depend on its algorithmic novelty and time and space complexity. However, the availability and implementation of the software have become more and more important (Nguyen et al., 2018). Since there are many tools available in the market, if a method is not well-implemented, potential users will simply pick another tool that is ready-to-run. It is unlikely that life scientists, who are the intended audience of these software, invest time to learn a programming language in order to implement complex algorithms reported in some papers. Practically, input and output format, graphical user interface, programming language, user-friendliness, and documentation are all important factors to be considered. More importantly, since reproducibility has become an outstanding issue recently, software availability and version control are critical for quality control and for reproducing the results reported in scientific papers (Sandve et al., 2013). For that reason, many journals today require authors to make their code available before publishing the paper.

At the time of this review, all of the 22 surveyed methods are available as either a standalone package or a web-based tool. Among these, there are 20 standalone packages and three web-based tools (one tool has both standalone package and web-based tool). Standalone tools are more often chosen to make use of the computational power of users' local machines or servers. Some of these packages provide a friendly interface for users to interact with. These software usually provides interactive features for users to manipulate the network and explore the data, which is illustrative and convenient. Some of them, e.g., PinnacleZ, BMRF-Net, and jActiveModules, are distributed as plugins of Cytoscape (Shannon et al., 2003) to

TABLE 1 | Computational tools for active subnetwork identification.

Method	Availability	Web	Pkg	Code	GUI	Cml	References	Cit.	Cit./year	Lic.
GREEDY ALGORITHM										
DIAMOnD	https://github.com/barabasilab/DIAMOnD/	✗	✓	Python	✗	✓	Ghiassian et al., 2015	74	24	**free
GXNA	http://statweb.stanford.edu/%7eserban/gxna/	✗	✓	C++	✗	✓	Nacu et al., 2007	141	12	free
MATISSE	http://acgt.cs.tau.ac.il/matisse/	✗	✓	Java	✓	✗	Ulitsky and Shamir, 2007	313	28	free
CEZANNE	http://acgt.cs.tau.ac.il/matisse	✗	✓	Java	✓	✗	Ulitsky and Shamir, 2009	111	12	free
PinnacleZ	http://apps.cytoscape.org/apps/pinnaclez	✗	✓	Java	✓	✗	Chuang et al., 2007	1,414	128	free
RME Module Detection	http://brl.bcm.tmc.edu/rme/index.rhtml	✗	✓	Ruby	✗	✓	Miller et al., 2011	82	11	MIT
EVOLUTIONARY ALGORITHM										
BMRF-Net	https://sourceforge.net/projects/bmrfcjava/	✗	✓	C++, Java	✓	✗	Shi et al., 2015	4	1	free
COSINE	https://cran.r-project.org/package=COSINE	✗	✓	R	✗	✓	Ma et al., 2011	64	9	GPL-3
GLADIATOR	http://www.cs.tau.ac.il/~roded/GLADIATOR.zip	✗	✓	Python	✗	✓	Silberberg et al., 2017	3	3	free
jActiveModules	http://apps.cytoscape.org/apps/jactivemodules	✗	✓	Java	✓	✗	Ideker et al., 2002	1,115	69	GPL
MOEA	https://github.com/WeiqiChen/MOEA-active-module-identification	✗	✓	MATLAB	✗	✓	Chen et al., 2017	1	1	free
DIFFUSION-FLOW EMULATION MODEL										
BioNet & Hienz	http://www.mi.fu-berlin.de/w/LiSA/Heinz	✗	✓	R, Python	✗	✓	Dittrich et al., 2008 Beisser et al., 2010	175 413	17 51	MIT
HotNet & HotNet2	http://compbio.cs.brown.edu/projects/hotnet2/	✗	✓	Python, MATLAB	✗	✓	Vandin et al., 2011 Leiserson et al., 2015	237 313	33 104	**free
RegMod	https://bmcbioinformatics.biomedcentral.com/articles/ 10.1186/1471-2105-11-26	✗	✓	MATLAB	✗	✓	Qiu et al., 2010	73	9	free
ResponseNet	http://netbio.bgu.ac.il/respnet/	✓	✗	Python	✓	✗	Lan et al., 2011 Basha et al., 2013	242 57 15	26 8 3	free
TimeXNet	http://timexnet.hgc.jp/	✓	✓	Java	✓	✓	Patil and Nakai, 2014	6	1	**free
RANDOM WALK ALGORITHM										
EnrichNet	http://lcsb-enrichnet.uni.lu/enrichnet/index.php	✓	✗	R, PHP	✓	✗	Glaab et al., 2012	171	28	**free
Walktrap-GM	https://github.com/petrochilos/walktrap-GM/	✗	✓	R	✗	✓	Petrochilos et al., 2013	20	4	GPL
MAXIMAL CLIQUE IDENTIFICATION										
MEMo	http://sanderlab.org/tools/memo.html	✗	✓	Java, Python	✗	✓	Ciriello et al., 2012	377	62	LGPL
ModuleDiscoverer	https://www.leibniz-hki.de/en/modulediscoverer.html	✗	✓	R	✗	✓	Vlaic et al., 2018	3	3	GPL
CLUSTERING-BASED METHOD										
ClustEx	http://bioinfo.au.tsinghua.edu.cn/member/jgu/clustex/	✗	✓	C++	✗	✓	Gu et al., 2010	54	6	free
SAMBA	http://acgt.cs.tau.ac.il/samba/	✗	✓	Java	✓	✓	Tanay et al., 2004	431	30	free

*Availability provides links to software. Web and Pkg indicate whether the web interface and the standalone package are available, respectively. Code lists the programming languages that were used to implement the software. GUI and Cml indicate whether the tool has GUI and command line, respectively. Reference provides the corresponding publication. Cit. shows the total number of citations while Cit/year shows the number of citations per year. License provides software license; GPL is an abbreviation for the GNU General Public License; LGPL stands for GNU Lesser General Public License; MIT stands for Massachusetts Institute of Technology; **free is free for academic and non-commercial use. Note that PinnacleZ is not compatible with Cytoscape 3 but it can still be downloaded from web archive https://web.archive.org/web/20120105141450/http://chianti.ucsd.edu/cyto_web/plugins/pluginjardownload.php?id=170.*

make use of its friendly interactive interface in manipulating networks. The rest provide command line interface or APIs for users to conduct experiments. Users usually need a third-party software to visualize the result networks such as Cytoscape. An advantage of tools with a command line interface is that it is easier for advanced users to integrate and embed these tools in their automated analysis pipeline. Most standalone tools require some administrative skills to install. Since these tools require interactome data, users are expected to download, locally store, and periodically update the network databases (partial or full copy). A standalone tool usually does not require internet access to perform analysis, which enhances the security and privacy of the experimental data.

Web-based tools (ResponseNet, TimeXNet, and EnrichNet), on the other hand, rely on a remote server to conduct analysis and provide computational power and a graphical interface; therefore, a local installation is generally not needed. Web-based tools are more user-friendly than standalone tools; however, they require an internet connection and a browser for access. In terms of cybersecurity and data privacy, this is considered a disadvantage compared to standalone tools. One major advantage of web-based tools is that most updates are transparent. In turn, this enhances the users' performance and enables collaboration between users by eliminating the burden of local installation and the need to keep it up-to-date.

The choice of the programming language used for the implementation also influences how well the method will be received. Tools that are well-implemented and packaged are more accepted than those that are poorly implemented or not user-friendly. Many methods implemented in Java provide good performance, can run on multiple platforms (Windows, Linux, MacOS), and offer a nice interactive user interface. For packages providing command line or APIs, it is worth to mention that the programming language plays a vital role in attracting users. For example, R users will prefer using an R package rather than learning a new language (such as Python or MATLAB). The programming language can also be an obstacle when there is a need to integrate a tool written in a different language to the current analysis pipeline. Most tools published as R packages can be easily installed due to R's user-friendly package manager. Other standalone tools written in C++, Python, and Ruby provide a command line interface to execute the analysis. Tools implemented in C++ also need to be compiled before using.

We also report the number of citations (and citations per year) for each method according to Google Scholar. Although the number of Google citation is not the right metric to assess a method's novelty or performance, it partially reflects how well a tool is accepted or known among researchers in the community. Finally, we report the license of each software. All of the surveyed software are free-of-charge for academic purposes. Many of them are freely available for non-academic users as well.

2.2. Experimental Data and Network Databases

Table 2 shows the input of each method, as well as the corresponding network databases and applicable species. Up to date, most methods are designed for analyzing human diseases using protein-protein interactions. Among the 22 methods, only six were designed to work with other species, including *Rattus norvegicus* (ModuleDiscoverer), *Mus musculus* (MATISSE, CEZANNE, TimeXNet), *Saccharomyces cerevisiae* (MATISSE, CEZANNE, jActiveModule, ResponseNet, TimeXNet, SAMBA), *Drosophila melanogaster* (MATISSE, CEZANNE), and *C. elegans* (MATISSE, CEZANNE). Most methods claim to be able to work with other species provided that the interaction network is available.

A subnetwork detection analysis typically requires two different kinds of input: (1) experimental data, and (2) interactome networks. Experimental data is generally data obtained from high-throughput technologies, such as gene and protein expression, somatic mutation, and copy number alteration. Among the 22 methods, only BioNet & Heinz uses the survival information to score genes in addition to differential analysis of expression data (**Supplementary section 1.12**). Most methods are designed for comparative analysis of two phenotypes, e.g., condition (disease) vs. control (healthy). Among the 22 methods, only four methods can detect subnetworks that are perturbed across multiple diseases or conditions. These are PinnacleZ, COSINE, GLADIATOR, and jActiveModules. These methods use statistics and tests (e.g., *F*-test, mutual information) that are able to compare more than two groups of samples in order to score the candidate subnetworks (see section 6 for discussions).

Different analysis methods use different input formats. There are only three methods that use mutation profiles as input, including RME Module Detection, HotNet, and MEMo. These methods aim to identify subnetworks that have more genes with mutations than expected. Most other methods accept the whole gene expression matrix, in which rows represent genes and columns represent samples from different phenotypes. Some methods accept only differentially expressed (DE) genes/proteins and their corresponding statistics (fold-change, *p*-value). TimeXNet is the only method that requires time-course data in the format of DE genes. The list of DE genes or proteins can be obtained by using a predefined cut-off based on *p*-value, fold-change, or both. Network approaches relying on input DE genes, however, might be overly sensitive to both selection method and cut-off threshold. First, a slight change in the threshold can greatly alter the list of DE genes (Nam and Kim, 2008). Second, different selection methods often produce different lists of DE genes. For example, the list of genes with high fold-changes is often not identical to the list of genes with significant *p*-values. In addition, for the same disease, independent studies or measurements often produce inconsistent sets of DE genes (Tan et al., 2003; Ein-Dor et al., 2005, 2006). This makes network analysis methods that rely on DE gene list even less reliable.

The input format for each method can be different depending on the programming language and the implementation of the method. For R packages, a common input of gene expression profiles is usually a matrix object where rows represent samples and columns represent genes (or vice versa). The input network can be passed as an adjacency matrix representing the relationship between nodes. For Cytoscape's plugins, the input network is in the format of Cytoscape network input files. With other methods, gene expression and network data are usually stored in flat files with a specific format defined by the software.

Besides experimental data, most methods also require interactome data or biological networks for their analysis. Biological networks are graphs representing either protein-protein networks or gene-gene networks. In these networks, the nodes are proteins or genes, and edges are interactions between them. Each network can contain additional information such as directions of the interactions (in directed networks), weights of nodes and edges, and other knowledge about the proteins, genes or their interactions (e.g., different types of interactions). RME Module Detection is the only method that does not require a predefined network as input. It constructs the global network from mutation data before extracting active subnetworks.

Among the network databases presented in **Table 2**, many of them are widely used in pathway analysis such as KEGG, HPRD, STRING, Reactome, NCI-PID, and BioGRID. In the subnetwork analysis, while most methods use a single protein network database to conduct the experiment, some methods, such as jActiveModules, MOEA, MEMo, TimeXNet, and DIAMOnD, combine multiple networks from different sources to construct a larger network. Since overlap among network databases is small (Chaurasia et al., 2006; Prieto and De Las Rivas, 2006), combining multiple databases can potentially increase the knowledge about interactome networks to build a more comprehensive biological network.

3. METHODS

Figure 1 shows the schematic representation of computational approaches that integrate phenotypic molecular profiles with known interactions accumulated in network databases. Most methods start by scoring the nodes and calculate node similarity that reflects the expression change (e.g., between disease and control) and the correlation between genes, respectively. Then, they adjust the scores and edge weights by taking into consideration the topological order and interaction between genes and proteins. The next step is to construct the subnetworks using edge weights and node scores. Typically, each method deploys its own subnetwork extension strategy in order to optimize a particular subnetwork scoring function using node scores and edge weights. After the subnetworks are constructed, each method performs a hypothesis testing to assess the statistical significance of identified modules. Some methods also repeatedly reconstruct the subnetwork after the statistical tests in order to find a more optimal solution.

Here we divide the methods into 6 categories according to the way the subnetworks are expanded: (1) greedy algorithms, (2) evolutionary algorithms, (3) maximal clique identification, (4) random walk algorithms, (5) diffusion emulation models, and (6) clustering-based methods. We summarize the methods in each category, providing the big picture and insights. Section 6 contains a detailed characteristics of each method.

3.1. Greedy Algorithms

In this section, we review six approaches that utilize a greedy strategy in order to construct active subnetworks, including GXNA (Gene eXpression Network Analysis), CEZANNE (Co-Expression Zone ANalysis using NEtworks), MATISSE (Module Analysis via Topology of Interactions and Similarity SEts), DIAMOnD (DIseAse MOdule Detection), PinnacleZ, and RME (recurrent and mutually exclusive) Module Detection.

The common flow of greedy algorithms consists of three major steps: (i) seed nodes selection (ii) subnetwork expansion, and (iii) significance testing. In the first step, the seeds can be randomly selected nodes (GXNA and PinnacleZ), high-scoring nodes (MATISSE and CEZANNE), user-defined nodes (DIAMOND and GXNA), or all nodes in the network (RME Module Detection). In the second step, each method then greedily extends the seeds with neighboring nodes with the objective to maximize the subnetwork's score. The procedure is repeated until further expansion does not increase the objective function. Some methods introduce early stopping criteria, such as the maximum size (RME Module Detection) or the improvement rate (PinnacleZ). In the third step, the statistical significance of the identified subnetworks is assessed by comparing its score against the scores obtained from random subnetworks (CEZANNE, PinnacleZ, RME Module Detection), or from permuting sample and gene labels (GXNA, MATISSE, PinnacleZ). This statistical significance of a subnetwork represents the probability of observing such score or higher, just by chance. The smaller the p-value, the less likely that such extreme score is observed by chance, i.e., the more likely the subnetwork has significant changes or significantly perturbed under the impact of the disease. DIAMOnD is the only method in this category that does not assess the statistical significance of the resulted subnetworks.

Greedy algorithms are fast and intuitive. However, since the decision at each step aims to improve the current state of the solution without paying attention to the global situation, it does not guarantee to produce the most optimal path. In fact, there is a high chance that the greedy algorithm does not find the global optima. Therefore, the selection of starting points plays a vital role in identifying optimal solutions. In addition, since this approach depends heavily on maximizing the score of the network by repeatedly adding adjacent nodes, the scoring function plays a vital role in the entire analysis, affecting the construction as well as the statistical significance of the obtained subnetworks. The methods scoring the network based on the similarity or correlation in gene expression change (MATISSE, CEZANNE, RME Module Detection, and PinnacleZ) tend to expand the modules to contain only highly similar genes, which can result in subnetworks missing important intermediate genes. Moreover, in some cases, these methods can produce large subnetworks with hundreds of genes that are difficult to interpret.

3.2. Evolutionary Algorithms

Here we review five approaches that use evolutionary algorithms to search for active modules with optimal scores: BMRF-Net (Bagging Markov Random Field), COSINE (COndition SpecIfic sub-NEtwork), GLADIATOR (GLobal Approach for DIsease AssociaTed mOdule Reconstruction), jActiveModule, and MOEA (Multi-Objective Evolutionary Algorithm). Similar to greedy approaches, evolutionary methods first define a scoring formula for each node and each edge as well as for a subnetwork whose score is often a weighted aggregation of nodes and edges belonging to the subnetwork. Each tool then uses either Generic

TABLE 2 | Active module identification approaches along with their corresponding input, network databases and species.

Method	Experimental input	MC	Network database	Species
GREEDY ALGORITHM				
DIAMOnD	Gene expression	✗	TRANSFAC, IntAct, MINT, BioGRID, HPRD, CORUM, PhosphositePlus, curated network (Vinayagam et al., 2011)	*Homo sapiens*
GXNA	Gene expression	✗	EntrezGene, KEGG	*Homo sapiens*
MATISSE & CEZANNE	Gene expression	✗	SGD, BioGRID, BIND, HPRD, GO, MIPS, KEGG	*Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*
PinnacleZ	Gene expression	✓	GO, Cell Circuits	*Homo sapiens*
RME Module Detection	Somatic mutation (SNP, CNV)	✗	–	*Homo sapiens*
EVOLUTIONARY ALGORITHM				
BMRF-Net	Gene expression	✗	HPRD	*Homo sapiens*
COSINE	Gene expression	✓	HPRD	Homo sapiens
GLADIATOR	Multiple lists of proteins where each list is associated with a disease	✓	Curated Network (Menche et al., 2015)	*Homo sapiens*
jActiveModules	Gene expression	✓	BIND, TRANSFAC, GAL	*Saccharomyces cerevisiae*
MOEA	Gene expression	✗	BioGRID, KEGG, GO	*Homo sapiens*
DIFFUSION-FLOW EMULATION MODEL				
BioNet & Hienz	Gene expression& survival information	✗	HPRD	*Homo sapiens*
HotNet & HotNet2	Mutation frequency of genes	✗	KEGG, HPRD	*Homo sapiens*
RegMod	Gene expression	✗	HPRD, MSigDB, GO	*Homo sapiens*
ResponseNet	Weighted list of DE genes and proteins	✗	Curated Network (Yeger-Lotem et al., 2009)	*Saccharomyces cerevisiae*
TimeXNet	Three list of DE genes and log fold change at initial intermediate and late stages	✗	HitPredict, InnateDB, TRANSFAC, KEGG	*Mus musculus*, *Saccharomyces cerevisiae*
RANDOM WALK ALGORITHM				
EnrichNet	A list of genes or proteins	✗	STRING, KEGG, BioCarta, WikiPathways, Reactome, InterPro, NCI-PID	*Homo sapiens*
Walktrap-GM	Gene expression	✗	HPRD, KEGG	*Homo sapiens*
MAXIMAL CLIQUE IDENTIFICATION				
MEMo	Gene expression, somatic mutation, copy number variation	✗	Reactome, MSKCC Cancer Cell Map, NCI-PID, HPRD, GO, PANTHER, INOH	*Homo sapiens*
ModuleDiscoverer	Gene expression	✗	STRING	*Rattus norvegicus*
CLUSTERING-BASED METHOD				
ClustEx	DE genes and fold change	✗	HPRD	*Homo sapiens*
SAMBA	Gene expression, transcription factor (TF) binding	✗	MIPS	*Saccharomyces cerevisiae*

Experimental Input indicates various input requirements that need to be supplied for the method to function. MC indicates whether the tool can handle multiple conditions (i.e., multiple diseases instead of the typical disease vs. control). Network Database provides network interaction information for various kinds of methods. Species gives information about species whose data could be used by the computational tools.

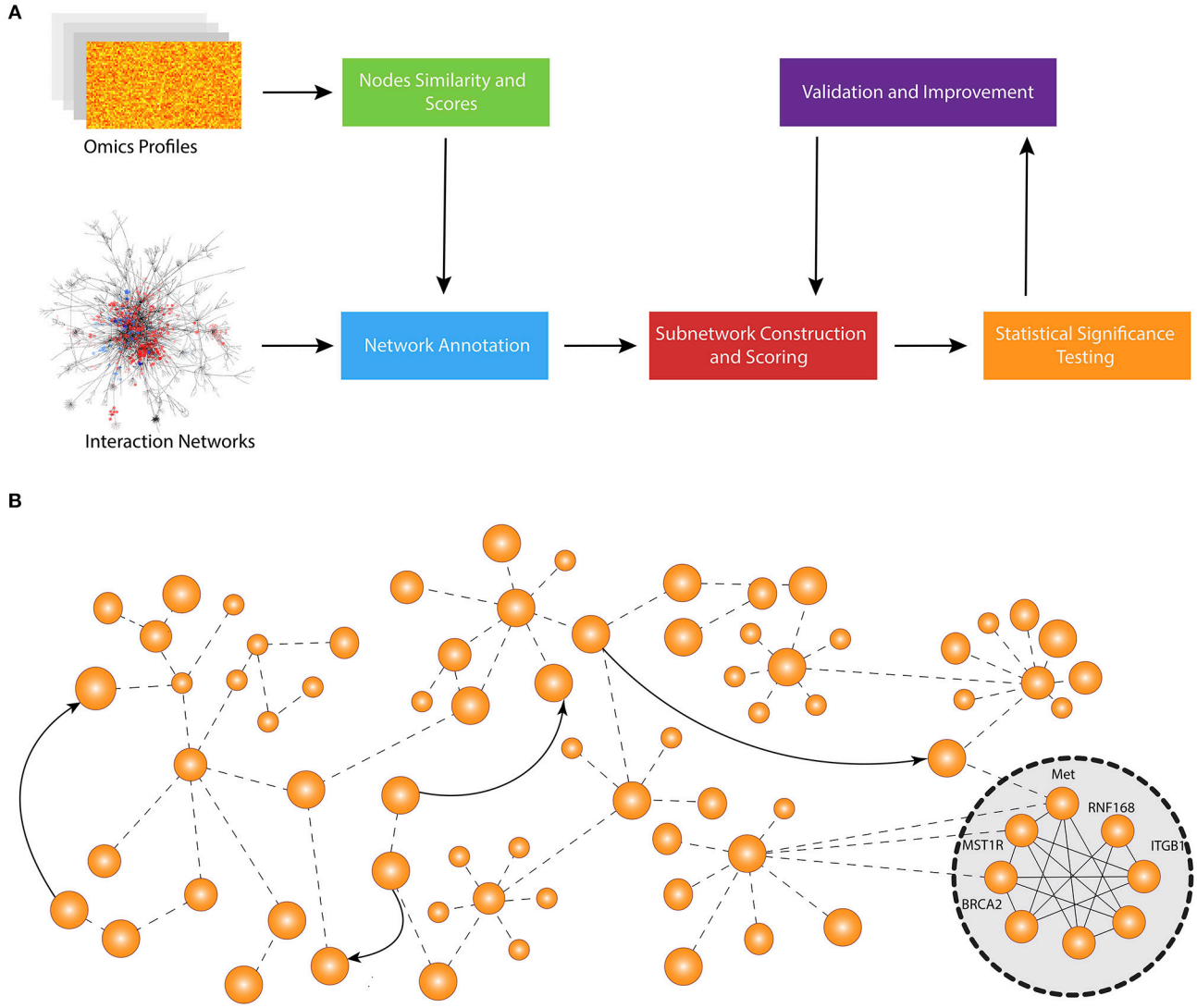


FIGURE 1 | Overall workflow of active subnetwork identification. **(A)** Schematic representation of computational approaches that integrate molecular profiles with known interactions accumulated in knowledge bases. Most methods start by scoring the nodes and calculating node similarity that reflects the expression change (e.g., between disease and control) and correlation between genes, respectively. Then, they adjust the scores and edge weights by taking into consideration the topological order and interaction between genes and proteins. The next step is to construct the subnetworks using edge weights and node scores. Typically, each method develops a specific subnetwork extension strategy in order to optimize a specific subnetwork scoring function using node scores and edge weights. After the subnetworks are constructed, each method performs a hypothesis testing to assess the statistical significance of identified modules. Some methods also repeatedly reconstruct the subnetwork after statistical tests to find a more optimal solution. **(B)** An example network and identified active subnetwork. The subnetwork are often a very well-connected component of the global network.

Algorithm (COSINE and MOEA) or Simulated Annealing (BMRF-Net, GLADIATOR, and jActiveModules) to search for an optimal subnetwork with the highest aggregate score. Among the five methods, only BMRF-Net and jActiveModule access the statistical significance of the obtained subnetwork using resampling and bootstrap, respectively.

Abstractly, the subnetwork construction can be formulated as a global optimization problem. Given p as the total number of genes, a subnetwork is represented as a binary vector of length p. The i^{th} element in the vector being 1 means that the i^{th} gene is present in the network. Evolutionary algorithms seek to find a binary vector that optimizes a certain scoring function. Simulated Annealing (SA) algorithm initializes a subnetwork by assigning each node as either active or inactive with a probability (default $\frac{1}{2}$). At each iteration, the algorithm randomly chooses a node and toggle the node's state (from active to inactive and vice versa). It then recalculates the aggregate score of the subnetwork. If the new score is greater than the old score, the state of the node is kept toggled. Otherwise, the node is kept toggled with a certain probability (to avoid being trapped in a local minimum). The algorithm returns the highest scoring subgraph after a number of iterations. Note that GLADIATOR maximizes the similarity

(using Jaccard index) between the connected modules provided for different diseases instead of optimizing the aggregate score of nodes and edges. The classical simulated annealing algorithm gets its inspiration from heat treatment in metallurgy which involves annealing metal to increase crystal size while reducing defects (Kirkpatrick et al., 1983).

Genetic Algorithms (SA), on the other hand, are inspired by natural selection, the process that drives biological evolution. The algorithm initialization sets certain genes (e.g., nodes with high scores) to 1 (active) and considers these genes as the starting population. Individuals in the population (parents) are then selected in pairs for reproduction based on their fitness score, in which crossover and mutation are happening. Crossover involves exchanging information from the parents to produce offspring while random mutations (with a low probability) alter the offspring to maintain diversity. The algorithm stops when the population has converged.

Although both GA and SA produce good quality solutions in the problem of finding optimal subnetworks, there is always a trade-off between running time and solution quality, which is affected by the size of the solution in GA and the temperature decay rate in SA (Adewole et al., 2012). The advantage of these algorithms is that they are not limited to the size and the complexity of the search space. Therefore, it can work with very large networks. In contrast to greedy algorithms, genetic algorithms aim to find the global solution and have proven to be very efficient in finding an approximation of global optima. Since GA and SA are both efficient in solving the problem of finding optimal subnetworks, it is important that the scoring process reflects precisely the perturbation and signal propagation of the subnetworks.

3.3. Diffusion-Flow Emulation Models

In this section, we discuss five methods that emulate diffusion flow phenomena in order to construct active subnetworks. Two of these are inspired by the heat diffusion process (HotNet and RegMod), while three others by the water flow phenomenon (BioNet & Heinz, ResponseNet, and TimeXNet). These are methods that aim to find a global solution through algorithmic optimization. Among the five, only TimeXNet and HotNet provide a statistical significance of the obtained active modules by using a permutation test.

Given a weighted and directed protein-protein interaction (PPI) network, BioNet & Heinz, ResponseNet, and TimeXNet emulate an abstract flow from a source node to a sink node through capacity- and cost-associated edges. The objective is to minimize the total cost from a source node to a sink node through a linear formulation in which variables represent the flow over each edge. Each edge of the network is assigned with: (i) a cost that is inversely proportional to the interaction reliability between the two connected nodes, and (ii) a flow capacity that is proportional to the similarity in molecular measurements of the two connected genes. The optimization problem is then solved using constrained linear programming, in which constraints (linear equalities or inequalities) are given to nodes and edges. While ResponseNet and TimeXNet produce only one optimal solution, Heinz allows users to explore different sub-optimal networks by adding a hamming distance to the optimal subnetwork to constrain the differences of the returned sub-optimal networks.

Heat diffusion algorithms, HotNet and RegMod, define the problem of finding active subnetworks as a heat diffusion model. Given a PPI network in which nodes weight represents initial heat, the optimization process delivers heat through edges until the heat in the network is equilibrium. Hot subnetworks are constructed after selecting edges transferring a total heat amount larger than a certain threshold. RegMod uses a heat diffusion kernel to calculate the similarity between two nodes, then computes the score for each gene that represents its relationship with other genes in the network. Active subnetworks are obtained by extracting high scoring genes.

3.4. Random Walk Algorithms

A random walk is a simulated path consisting of successive random transitions through a mathematical space, for example, an integer set or a 2-D plane. The transitions are not necessarily a complete random action but rather can be biased toward a specific direction. In a biological network, the connections (or interaction intensities) between different pairs of proteins are different. When applying the random walk algorithm on the network, the walk is more likely to stay in the subnetwork with high interactions among the members, because the chance of the walk choose the lower interaction paths to escape the subnetwork is small. The performance of the algorithm is heavily affected by the method used to weight the interactions.

Walktrap-GM (R package) and EnrichNet (web interface) are the two tools that utilize random walks to identify active subnetworks. EnrichNet requires a list of starting proteins while Walktrap-GM uses as input gene expression data. To build the weighted interaction network, Walktrap-GM calculates the weight of each edge as the average fold-change of the two connected nodes. In contrast, ErinchNet uses the weight extracted from STRING 9.0 database, which could be argued to be better, as it is combined confidence from different sources such as curated databases, gene co-occurrence, text mining, etc. To travel in the network, Walktrap-GM uses the random walk algorithm which transits from the current node to its adjacent nodes with a probability based on the weight of the linking edge and the degree of the current node. Using the transition probabilities, the distance between two nodes and between two communities formed by the random walk process are then calculated. The traverse will merge two communities if it minimizes the mean of the squared distances between each node and its community. On the other hand, EnrichNet uses a random walk with restart to emphasize the importance of starting nodes. Therefore, the result would be subnetworks that has strong connections with the input gene list. Overall, Walktrap-GM is expected to be more useful to look for new active modules while EnrichNet is expected to perform better in the deeper investigation of already-known modules. Walktrap-GM also assesses the statistical significance of the obtained subnetworks using bootstrap.

3.5. Maximal Cliques Identification

This class of methods for active subnetwork identification is focusing on finding cliques, i.e., subnetworks in which every node is connected with all others. This approach is based on the assumption that all the proteins in an active module would have tight connections with the rest. Due to the lack of efficient algorithms to find these cliques in a dense network (Tomita et al., 2006), a preprocessing step to simplify the network is necessary. Two methods in this review (MEMo and ModuleDiscover) have different solutions to this problem. ModuleDiscover tries to filter out the interactions that are not strong enough based on the data from STRING database. In contrast, MEMo applies three different kinds of filters based on significantly mutated gene, copy number regions of interest and mRNA expression to retain only the altered genes in their network. Then, cliques are extracted from these filtered networks.

The advantage of these methods is the reliability of the identified subnetworks, due to the nature of clique (strong interaction in subnetwork). Moreover, by modifying the algorithm, various kinds of data could be applied to the simplification step to refine the network even more before the identification of active modules. As a potential disadvantage, ModuleDiscover's reliance on the prior knowledge in the STRING database means that the discovery of new modules is essentially impossible. Also for MEMo, the aggressive filtering (three filter layers) means that some important information may be lost in the process.

3.6. Clustering-Based Methods

In this section, we review two methods using different approaches in the identification of active modules from other groups. These are ClustEx, which is based on a hierarchical clustering algorithm, and SAMBA, which uses biclustering on a bipartite graph. ClustEx first calculates weights and distances for the edges using the Pearson correlation of the expression of the genes associated with the nodes. Subsequently, ClustEx clusters the genes using hierarchical clustering. It then identifies the active modules through two steps. In the first step, it looks for node pairs with the distance below a given threshold and assigns the connecting path as the initial clusters. In the second step, it expands the initial cluster to the surrounding genes. Finally, the nodes that are visited by the 10-shortest path in this expanded cluster are identified as belonging to an active module. As potential limitations, one can note that during the first step, ClustEx calculates the distances between every pair of nodes which could be a heavy computational task. Moreover, due to the nature of the expanding process, which is determined by the 10-shortest path, some important nodes in a tightly connected cluster could be left out.

Unlike other methods, SAMBA takes a completely different approach to identify the active modules. Instead of building one single interaction network using genes as nodes, they build a weighted bipartite graph where nodes on one side represent the genes and nodes on the other side represent properties of proteins encoded by them. The connection between two parts represents the probability for a gene to have a specific property. The locally optimal subgraphs are identified using biclustering and overlapping is minimized by limiting the shared properties between subgraphs. The performance of this model is heavily dependent on the selections of properties layer, which could make it challenging to apply SAMBA to a new disease.

4. APPLICATIONS

The 22 surveyed methods have been widely applied in real-world scenarios to find disease gene signatures, dysfunctional pathways, common mechanisms of different diseases, as well as to discover drug and toxicity effects on different organisms. PinnacleZ, despite being the most highly cited method, was mostly cited for the discovery reported in the paper. The method jActiveModules appears to be the most used tool for discovery and understanding biological mechanisms. As a Cytoscape plugin, jActiveModules has its advantages in network visualization and manipulation. At the time of this survey, we found approximately 80 studies that utilized this software. BioNET as an R package, was also applied in real-world settings in more than 30 studies. Other tools including EnrichNet, MEMo, and MATISSE were utilized in more than 10 studies. The number of studies and manuscript DOIs are reported in **Supplementary section 2**.

PPI networks have been widely used in identifying disease biomarker and sample classification. For example, Chuang et al. (2007) used PinnacleZ to classify patients with breast cancer and Yuan et al. (2017) applied jActiveModules to find gene signatures for leukemia patients for sample stratification purposes. Network-based signatures have proven to be more reliable and reproducible than signatures identified from gene expression data alone. In fact, proteins involved in cancers tend to show a high level of connectivity in the PPI networks (Jonsson and Bates, 2006). Therefore, discovering gene signatures for specific diseases can be greatly improved by identifying significantly impacted subnetworks from the PPI network, especially when the known disease genes are highlighted as seed genes in the network (Shafi et al., 2019).

Active subnetwork approaches have also been utilized to discover dysfunctional pathways of diseases. For example, Skov et al. (2012) used jActiveModules to study biological pathways and networks that are dysregulated in type 2 diabetes. The study by Riazuddin et al. (2017) made use of MATISSE to find novel gene candidates for the biological pathways of intellectual disability disorder. Similarly, Sharma et al. (2015) identified key pathways within the asthma module discovered by the DIAMOnD method. Resulted modules from subnetwork methods provide meaningful insights to dysfunctional processes of the underlying disease phenotypes.

The discovered subnetworks, although are resulted from a specific disease, can also be used to predict disease-causing genes for similar diseases or other complex diseases (Oti and Brunner, 2007). By identifying responsive modules corresponding to certain diseases, the associations among diseases can be discovered by network similarity. For example, Dong et al. (2014) used jActiveModules to extract active modules of type 2 diabetes and coronary heart disease to find pathways that are important to both diseases. In another study, Wuchty et al.

(2010) used significant subnetworks discovered by PinnacleZ to find pathways that help to discriminate major glioma subtypes. Studying associations between dysfunctional modules from different phenotypes may reveal the true mechanism of complex diseases such as cancers.

Drug and toxicity studies have also made use of active subnetwork identification to discover their effects on different organisms. For example, jActiveModules was used to identify the network regions that is active under methamphetamine (Bortell et al., 2017) and dioxin (Alexeyenko et al., 2010) exposure. Similarly, BioNet was applied to extract top-scoring networks to understand the impact of drug combinations on lymphoma disease (Zhao et al., 2014). BioNet was also used to find candidates for drug targets (Cursons et al., 2015). Although drug targets are typically regarded as single proteins, most drugs often interact with a larger number of proteins. By studying drug-response subnetworks, the overall effects of a drug can be revealed not only the efficacy of the drug to the target proteins but also its side effects.

5. CHALLENGES IN SUBNETWORK IDENTIFICATION

Even though subnetwork identification methods have been applied in many real-world applications, there are challenges that have not been addressed. In this section, we highlight the limitations of existing knowledge bases, as well as identify outstanding challenges from the method development perspectives.

5.1. Knowledge Bases

One major challenge is that most PPI knowledge bases are incomplete. For example, widely used networks in HPRD and BioGRID cover at most 50% of the known human PPIs (De Las Rivas and Fontanillo, 2010). In consequences, analysis results using these knowledge bases may be incomplete due to possible omissions of important factors. Another example is that the number of genes in KEGG remained around 5,000 over the past few years whereas the number of protein-coding genes is estimated to be between 19,000 and 20,000 (Ezkurdia et al., 2014). Integrative methods using networks and gene expression data are forced to work on a much reduced gene space. In many cases, using the reduced number of features in a classification algorithm decreases the classification performance (Staiger et al., 2013), suggesting that some important features (genes) had been left out by the PPI networks. One approach to increase the coverage of the PPI networks is to combine multiple knowledge bases to build a more comprehensive biological network.

Another important limitation of existing knowledge bases is that they are unable to keep up with the high-resolution information that has become available with the advancement of high-throughput technologies and multi-omics assays. For example, data obtained from RNA-Seq experiments allows us to identify transcripts that are active under certain conditions. Multiple transcripts mapping to the same gene can have distinct or even opposite functions due to the alternative splicing (Wang et al., 2008). Although this information is crucial to reveal the underlying biological mechanisms, the majority of the PPI knowledge bases only provides information at the gene level. In addition, knowledge bases do not provide information regarding cell types, conditions and time points each of which is essential to reveal the true phenomenon of a given biological condition. Finally, existing knowledge bases offer at most limited options to integrate multi-omics data. In the past decade, molecular data of all kinds, from transcriptomics to genomics, epigenetics, and non-coding RNA have accumulated on public repositories with unprecedented rates. However, most subnetwork approaches are limited to gene/protein data. A great wealth of these data remain unused since knowledge bases mostly store information about protein or gene interactions. Future approaches need to develop graphical models that take into consideration changes at different levels (e.g., methylation, miRNA, mRNA) to exploit the complementary information available in different types of omics data.

5.2. Method Development

One key challenge for subnetwork method development is the lack of universally accepted gold-standard to validate the identified subnetwork modules. Computational approaches are typically assessed by simulated data or by well-studied biological datasets (He et al., 2017; Vlaic et al., 2018). The advantage of using simulation data is that the ground truth is always known. Thus it can be used to compare different methods using sensitivity and specificity. However, simulation is often oversimplified and unable to capture the complexity of living organisms. On the other hand, when using real biological data, the biology is never fully known. In addition, many papers presenting new analysis methods include results obtained from only a couple of datasets and researchers are often influenced by the observer-expectancy effect (Sackett, 1979). Designing benchmark real datasets where the true mechanisms are known would help address this issue.

Furthermore, the majority of the active subnetwork identification methods do not account for the complexity of protein interactions. Most of active subnetwork identification methods and network clustering approaches, produce only non-overlap modules. These methods were developed based on the assumption that a protein can be active in at most one module. However, it is known that most proteins may involve in many biological processes. In addition, a disease can go through different stages and a protein may take place in different active modules at different time points. Producing large networks containing all possible interactions is insufficient to reveal the underlying mechanisms for complex diseases. Reporting different networks for different stages of dynamic networks, in this case, will significantly help to interpret the signal of the disease.

Finally, *p*-value-based approaches are subject to potential bias under the null hypothesis. In principle, the null distribution is used to assess the significance of the observed result obtained from an experiment. The *p*-values obtained from any sound statistical test are assumed to follow a uniform distribution (with the interval 0–1) (Fodor et al., 2007; Barton et al., 2013). Although this issue is yet to be investigated in the field

of subnetwork identification, it has been shown that many computational methods for network and pathway analysis have a systematic bias toward pathways related to cancer and well-studied diseases (Nguyen et al., 2017). In the study, the authors created a large pool of healthy individuals and then randomly compare two groups of healthy people. Interestingly, the *p*-value distributions of cancer pathways are extremely biased toward zero and thus are found significant in many analyses that have nothing to do with cancer. Similarly, subnetwork analyses are expected to be biased toward well-studied diseases and network modules. To overcome this problem, *p*-values of the candidate modules should be calculated under the null to demonstrate that a method is not biased under the null.

6. DISCUSSIONS

All of the methods surveyed here aim to identify one or several active subnetworks for one or several input datasets. However, they differ in their assumptions about the relationship among genes, protein, or both, which leads to different scoring functions and traversal strategies. **Figure 2** shows the workflows of the 22 methods, highlighting their characteristics and differences. From left to right are the techniques applied in each approaches: (i) node scoring, (ii) edge scoring, (iii) algorithm used to construct the subnetworks, and (iv) statistical test for assessing the significance of the identified active subnetworks. Note that GLADIATOR does not score nodes nor edges but rather it uses the Jaccard Index between input gene sets (of different diseases) as the objective for its simulated annealing algorithm. In this review, we categorize the 22 approaches according to the way they construct their network (main algorithm).

The problem of finding optimal subnetworks with the highest network score is NP-hard. Therefore, many methods address this problem via a heuristic approach that does not guarantee a global optimum. Random walk and greedy algorithms construct their modules by initializing the seeds and greedily extend the modules. Therefore, the results obtained with these methods will depend on the choice of the seeds. In a large network of tens of thousands nodes, it is harder to find a seed that leads to the global optima. Diffusion-flow emulation models, on the other hand, model the problem as a mathematical optimization that aims to find the global optimum using algorithmic optimization. For example, maximum-flow algorithms assign flow capacity and flow cost to nodes and edges and then find the global optimum using constrained linear programming. These mathematical approaches guarantee to reach a global optimum. Similarly, evolutionary algorithms also aim to find the global optimum or at least an approximation of it. The algorithms allow for transitions to states with a lower score in order to avoid being trapped at a local maximum/minimum. In principle, with a large number of iterations, these algorithms are likely to find a global solution.

Maximal clique approaches and clustering-based methods are distinct from the rest in terms of their goals and objectives. Maximal clique methods do not aim to find connected nodes with the best score. Instead, they aim to find groups of genes that interact with one another (every pair of nodes in a clique has an edge between them). However, it is not necessarily true that all genes in a clique always take part in certain biological processes. In addition, these methods may miss intermediate genes or proteins that play important roles in connecting those cliques. The clustering-based methods, on the other hand, assume that co-expressed genes are all involved in the same cellular process (ClustEx) or there is a hierarchical structure in the biological network (SAMBA). Since clustering approaches aim to group high-similarity genes into the same cluster without paying attention to the size of each cluster, the output can be highly imbalanced, including extremely large subnetworks that fail to provide insights into the underlying mechanisms of the given phenotypes.

The methods surveyed here use a wide range of scoring functions to score the nodes and edges. Most of them (except GLADIATOR) provide a scoring function for nodes or edges, but only some of them take into account the scores of both nodes and edges. While node-based scoring approaches look at the significance of one gene or protein in the context of the whole network, edge-based scoring networks look at the strength of the relationship among protein or gene. Without paying attention to the weight of the relationship between proteins or genes, node-based scoring methods may produce subnetworks that have high scoring nodes but do not have a meaningful relationship between nodes. Also, the edge-based scoring network may produce subnetworks that contain highly similar genes but have low significance in the network. These resulted subnetworks, unfortunately, will be difficult to interpret. Methods that take into account the scores of both nodes and edges are likely to produce a more accurate active module.

Node and edge scoring functions are the building blocks of the subnetwork score. In principle, they should be the summary statistics that capture the network perturbation, signal propagation as well as the changes between different phenotypes. Since each test and score is based on a certain assumption, users need to check if the assumption of each test matches the property of their data. For instance, the *z*-test and *t*-test assume that the data follows a normal distribution, while methods using fold-change assume that the effect size is the most important factor to capture the difference between the two conditions. Note that fold-change is highly dependent on the background signal, i.e., a shift in the range will significantly change fold change (e.g., 101 compared to 103 vs. 1 compared to 3) (Drăghici, 2011). Furthermore, the *t*-test, *z*-test, and fold-change approaches can only compare two conditions, while the F-test, mutual information, and binomial test allow users to capture changes across multiple conditions. Some methods do not take into account the scores of the nodes but rather require the user to input a gene list or protein list (GLADIATOR, ResponseNet, EnrichNet, and ClustEx) as the significant gene set to serve as starting points of the algorithm. These methods can be sensitive to the input gene list, in which small changes in the list can dramatically affect the resulted subnetworks. In contrast to the variety of node scoring, the edge scoring functions mostly rely on the

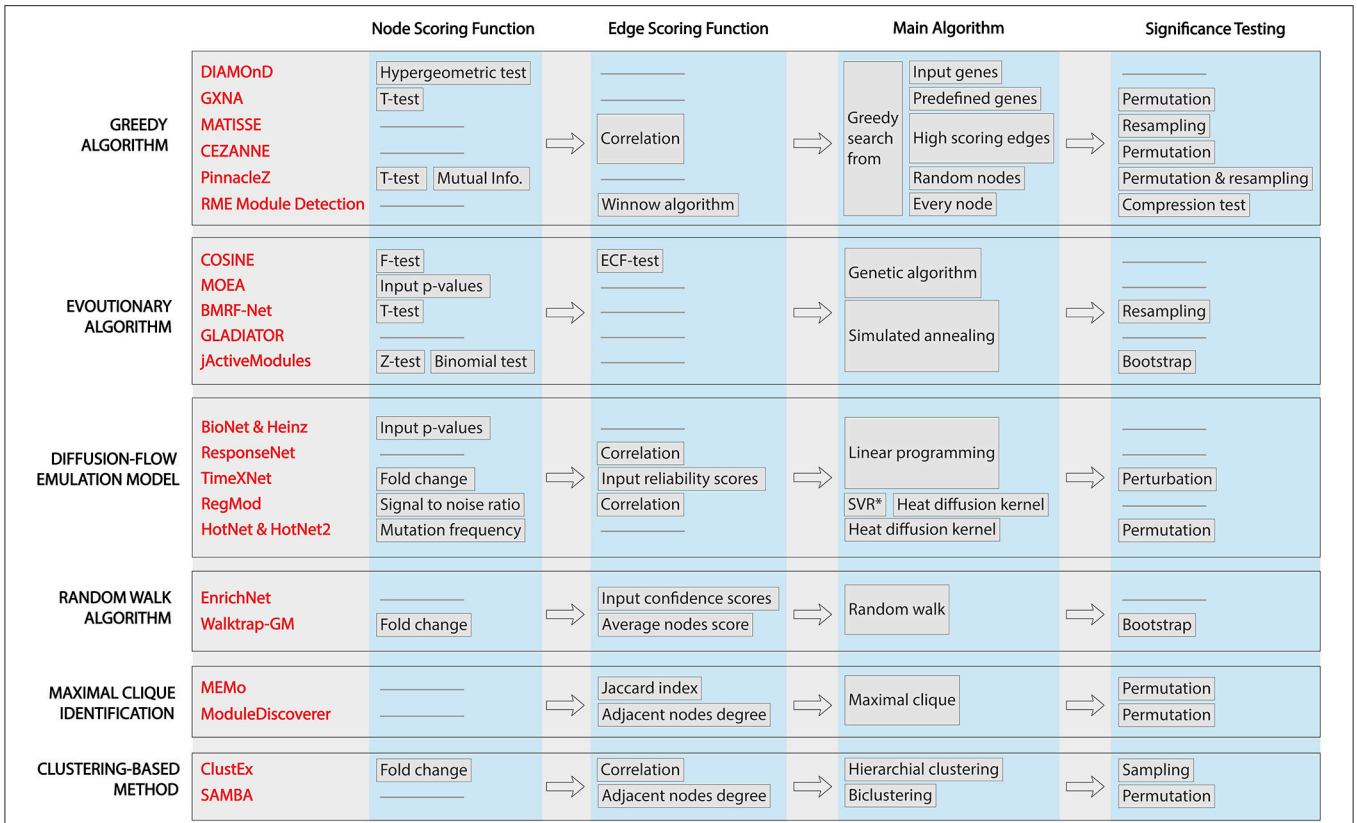


FIGURE 2 | Workflows of active module identification approaches. The figure highlights the key characteristics and key differences between each method. From left to right are the techniques applied in each approach: (i) node scoring, (ii) edge scoring, (iii) algorithm used to construct the subnetworks, and (iv) statistical test for assessing the significance of the identified active subnetworks. We note that GLADIATOR does not score nodes nor edges but uses Jaccard Index between input gene sets (of different diseases) as the objective for its simulated annealing algorithm.

correlation between two adjacent nodes to indicate the similarity between nodes.

The significance test used in a particular tool is also an important factor to consider. An aggregated score calculated for a module represents the level of signal perturbation or the degree of change observed in the subnetwork between different conditions. Similar to fold-change or effect-size, this score can be either the result of real biological changes or just by chance due to random noise. One needs to assess whether the observed change represents real biological differences. Therefore, a significance assessment should be done to assess how likely the aggregated score is observed just by chance under the null hypothesis, i.e., due to noise and chance alone. DIAMOnD, COSINE, MOEA, Bionet & Heinz, ResponseNet, and EnrichNet output the subnetworks and aggregated scores without performing a significance assessment. Thus, it is totally up to users to interpret the identified subnetworks and their scores. The remaining methods perform a significance assessment and calculate a *p*-value for each resulted subnetwork. For methods that provide multiple active subnetworks, a correction for multiple comparisons should be performed. Users can determine whether each subnetwork is significantly impacted by comparing the *p*-values with a pre-defined threshold.

7. CONCLUSIONS

In the past decades, there have been great efforts to mine network databases for identifying condition-specific cellular processes. One successful strategy has been to integrate these networks with molecular data to identify active subnetworks or modules that are involved in condition-specific biological functions. In this article, we review 22 methods that identify active subnetworks by integrating molecular data (e.g., expression profiles, protein, mutation) with known biological interaction accumulated in knowledge bases and public repositories. At the time of preparing this article, all surveyed methods are available as either a working standalone package or through a web-based interface. We categorize the 22 methods into five different categories according to the way they construct and extend the subnetwork. We summarize the pros and cons of each approach and category, focusing on their distinguishing characteristics and mathematical models. Our main objective is to help potential users and life scientists to choose methods that are most suitable for their available data and analysis purpose. This review will also help computational scientists to identify shortcomings of existing approaches in order to develop new methods that address current limitations.

AUTHOR CONTRIBUTIONS

TN and HN drafted the text with contributions from all co-authors: SS, DT, AS, and SD. SD reviewed and edited the final version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00155/full#supplementary-material

REFERENCES

Adewole, A., Otubamowo, K., Egunjobi, T., and Ng, K. (2012). A comparative study of simulated annealing and genetic algorithm for solving the travelling salesman problem. *Int. J. Appl. Inform. Syst.* 4, 6–12. doi: 10.5120/ijais12-450678

Alexeyenko, A., Wassenberg, D. M., Lobenhofer, E. K., Yen, J., Linney, E., Sonnhammer, E. L., et al. (2010). Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity. *PLoS ONE* 5:e10465. doi: 10.1371/journal.pone.0010465

Bapat, S. A., Krishnan, A., Ghanate, A. D., Kusumbe, A. P., and Kalra, R. S. (2010). Gene expression: protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer. *Cancer Res.* 70, 0008–5472. doi: 10.1158/0008-5472.CAN-10-0447

Barton, S. J., Crozier, S. R., Lillycrop, K. A., Godfrey, K. M., and Inskip, H. M. (2013). Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics* 14:161. doi: 10.1186/1471-2164-14-161

Basha, O., Tirman, S., Eluk, A., and Yeger-Lotem, E. (2013). ResponseNet2.0: revealing signaling and regulatory pathways connecting your proteins and genes–now with human data. *Nucleic Acids Res.* 41, W198–W203. doi: 10.1093/nar/gkt532

Beisser, D., Klau, G. W., Dandekar, T., Muller, T., and Dittrich, M. T. (2010). BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics* 26, 1129–1130. doi: 10.1093/bioinformatics/btq089

Bortell, N., Basova, L., Semenova, S., Fox, H. S., Ravasi, T., and Marcondes, M. C. G. (2017). Astrocyte-specific overexpressed gene signatures in response to methamphetamine exposure *in vitro*. *J. Neuroinflammation* 14:49. doi: 10.1186/s12974-017-0825-6

Chaurasia, G., Iqbal, Y., Hänig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. (2006). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* 35(Suppl. 1), D590–D594. doi: 10.1093/nar/gkl817

Chen, W., Liu, J., and He, S. (2017). Prior knowledge guided active modules identification: an integrated multi-objective approach. *BMC Syst. Biol.* 11:8. doi: 10.1186/s12918-017-0388-2

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562. doi: 10.1038/nrg.2017.38

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkv1351

Cursons, J., Leuchowius, K. J., Waltham, M., Tomaskovic-Crook, E., Foroutan, M., Bracken, C. P., et al. (2015). Stimulus-dependent differences in signalling regulate epithelial-mesenchymal plasticity and change the effects of drugs in breast cancer cell lines. *Cell Commun. Signal.* 13:26. doi: 10.1186/s12964-015-0106-x

De Las Rivas, J., and Fontanillo, C. (2010). Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* 6:e1000807. doi: 10.1371/journal.pcbi.1000807

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231. doi: 10.1093/bioinformatics/btn161

Dong, C., Tang, L., Liu, Z., Bu, S., Liu, Q., Wang, Q., et al. (2014). Landscape of the relationship between type 2 diabetes and coronary heart disease through an integrated gene network analysis. *Gene* 539, 30–36. doi: 10.1016/j.gene.2014.02.001

Drăghici, S. (2011). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. London, UK: Chapman and Hall; CRC Press.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21, 171–178. doi: 10.1093/bioinformatics/bth469

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5923–5928. doi: 10.1073/pnas.0601231103

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., et al. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* 23, 5866–5878. doi: 10.1093/hmg/ddu309

Fodor, A. A., Tickle, T. L., and Richardson, C. (2007). Towards the uniform distribution of null P values on Affymetrix microarrays. *Genome Biol.* 8:R69. doi: 10.1186/gb-2007-8-5-r69

Ghiassian, S. D., Menche, J., and Barabasi, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 28, i451–i457. doi: 10.1093/bioinformatics/bts389

Gu, J., Chen, Y., Li, S., and Li, Y. (2010). Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis. *BMC Syst. Biol.* 4:47. doi: 10.1186/1752-0509-4-47

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99. doi: 10.1038/nature02800

He, H., Lin, D., Zhang, J., Wang, Y.-p., and Deng, H.-w. (2017). Comparison of statistical methods for subnetwork detection in the integration of gene expression and protein interaction network. *BMC Bioinformatics* 18:149. doi: 10.1186/s12859-017-1567-2

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics* 18, S233–S240. doi: 10.1093/bioinformatics/18.suppl_1.S233

Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297. doi: 10.1093/bioinformatics/btl390

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2008). Human protein reference database–2009 update. *Nucleic Acids Res.* 37(Suppl. 1), D767–D772. doi: 10.1093/nar/gkn892

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Lan, A., Smoly, I. Y., Rapaport, G., Lindquist, S., Fraenkel, E., and Yeger-Lotem, E. (2011). ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res.* 39(Suppl. 2), W424–W429. doi: 10.1093/nar/gkr359

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 27, 1290–1298. doi: 10.1093/bioinformatics/btr136

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D., and Milosavljevic, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* 4:34. doi: 10.1186/1755-8794-4-34

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732. doi: 10.1038/nrg3552

Nacu, Ş., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics* 23, 850–858. doi: 10.1093/bioinformatics/btm019

Nam, D., and Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Brief. Bioinform.* 9, 189–197. doi: 10.1093/bib/bbn001

Nguyen, T., Mitrea, C., and Draghici, S. (2018). Network-based approaches for pathway level analysis. *Curr. Protoc. Bioinformatics* 61, 8–25. doi: 10.1002/cpbi.42

Nguyen, T., Mitrea, C., Tagett, R., and Draghici, S. (2017). DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis. *Proc. IEEE* 105, 496–515. doi: 10.1109/JPROC.2015.2507119

Oti, M., and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin. Genet.* 71, 1–11. doi: 10.1111/j.1399-0004.2006.00708.x

Patil, A., and Nakai, K. (2014). TimeXNet: identifying active gene sub-networks using time-course gene expression profiles. *BMC Syst. Biol.* 8:S2. doi: 10.1186/1752-0509-8-S4-S2

Petrochilos, D., Shojaie, A., Gennari, J., and Abernethy, N. (2013). Using random walks to identify cancer-associated modules in expression data. *BioData Mining* 6:17. doi: 10.1186/1756-0381-6-17

Prieto, C., and De Las Rivas, J. (2006). APID: agile protein interaction DataAnalyzer. *Nucleic Acids Res.* 34(Suppl. 2), W298–W302. doi: 10.1093/nar/gkl128

Qiu, Y. Q., Zhang, S., Zhang, X. S., and Chen, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* 11:26. doi: 10.1186/1471-2105-11-26

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044

Riazuddin, S., Hussain, M., Razzaq, A., Iqbal, Z., Shahzad, M., Polla, D., et al. (2017). Exome sequencing of pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability. *Mol. Psychiatry* 22, 1604–1614. doi: 10.1038/mp.2016.109

Sackett, D. L. (1979). Bias in analytic research. *J. Chron. Dis.* 32, 51–63.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32(Suppl. 1), D449–D451. doi: 10.1093/nar/gkh086

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* 9:e1003285. doi: 10.1371/journal.pcbi.1003285

Shafi, A., Nguyen, T., Peyvandipour, A., Nguyen, H., Draghici, S. (2019). A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Front. Genet.* 10:159. doi: 10.3389/fgene.2019.00159

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* 24, 3005–3020. doi: 10.1093/hmg/ddv001

Shi, X., Barnes, R. O., Chen, L., Shajahan-Haq, A. N., Hilakivi-Clarke, L., Clarke, R., et al. (2015). BMRF-Net: a software tool for identification of protein interaction subnetworks by a bagging Markov random field-based method. *Bioinformatics* 31, 2412–2414. doi: 10.1093/bioinformatics/btv137

Silberberg, Y., Kupiec, M., and Sharan, R. (2017). GLADIATOR: a global approach for elucidating disease modules. *Genome Med.* 9:48. doi: 10.1186/s13073-017-0435-z

Skov, V., Knudsen, S., Olesen, M., Hansen, M. L., and Rasmussen, L. M. (2012). Global gene expression profiling displays a network of dysregulated genes in non-atherosclerotic arterial tissue from patients with type 2 diabetes. *Cardiovasc. Diabetol.* 11:15. doi: 10.1186/1475-2840-11-15

Spirin, V., and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 12123–12128. doi: 10.1073/pnas.2032324100

Staiger, C., Cadot, S., Györffy, B., Wessels, L. F., and Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.* 4:289. doi: 10.3389/fgene.2013.00289

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39(Suppl. 1), D561–D568. doi: 10.1093/nar/gkq973

Tan, P. K., Downey, T. J., Spitznagel E. L. Jr, Xu, P., Fu, D., Dimitrov, D. S., et al. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31, 5676–5684.

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2981–2986. doi: 10.1073/pnas.0308661100

Tomita, E., Tanaka, A., and Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci.* 363, 28–42. doi: 10.1016/j.tcs.2006.06.015

Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* 1:8. doi: 10.1186/1752-0509-1-8

Ulitsky, I., and Shamir, R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics* 25, 1158–1164. doi: 10.1093/bioinformatics/btp118

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4:rs8. doi: 10.1126/scisignal.2001699

Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., et al. (2018). ModuleDiscoverer: identification of regulatory modules in protein-protein interaction networks. *Sci. Rep.* 8:433. doi: 10.1038/s41598-017-18370-2

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Wuchty, S., Zhang, A., Walling, J., Ahn, S., Li, A., Quezado, M., et al. (2010). Gene pathways and subnetworks distinguish between major glioma subtypes and elucidate potential underlying biology. *J. Biomed. Inform.* 43, 945–952. doi: 10.1016/j.jbi.2010.08.011

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* 41:316. doi: 10.1038/ng.337

Yi, S., Lin, S., Li, Y., Zhao, W., Mills, G. B., and Sahni, N. (2017). Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.* 18, 395–410. doi: 10.1038/nrg.2017.8

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104–110. doi: 10.1126/science.1158684

Yuan, X., Chen, J., Lin, Y., Li, Y., Xu, L., Chen, L., et al. (2017). Network biomarkers constructed from gene expression and protein-protein interaction data for accurate prediction of leukemia. *J. Cancer* 8, 278–286. doi: 10.7150/jca.17302

Zhang, K. X., and Ouellette, B. F. (2011). CAERUS: predicting cancer outcomes using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput. Biol.* 7:e1001114. doi: 10.1371/journal.pcbi.1001114

Zhao, J., Zhang, X. S., and Zhang, S. (2014). Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs. *CPT Pharmacometrics Syst. Pharmacol.* 3:e102. doi: 10.1038/psp.2013.79

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.

Copyright © 2019 Nguyen, Shrestha, Tran, Shafi, Draghici and Nguyen. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 10 April 2019
doi: 10.3389/fgene.2019.00265

NoMAS: A Computational Approach to Find Mutated Subnetworks Associated With Survival in Genome-Wide Cancer Studies

Federico Altieri[1], *Tommy V. Hansen*[2] *and Fabio Vandin*[1]*

[1] Department of Information Engineering, University of Padova, Padova, Italy, [2] Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council (CNR), Italy

Reviewed by:
Georges Nemer,
American University of Beirut, Lebanon
Salvatore Alaimo,
University of Catania, Italy

***Correspondence:**
Fabio Vandin
vandinfa@dei.unipd.it

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: 15 November 2018
Accepted: 08 March 2019
Published: 10 April 2019

Citation:
Altieri F, Hansen TV and Vandin F (2019) NoMAS: A Computational Approach to Find Mutated Subnetworks Associated With Survival in Genome-Wide Cancer Studies. Front. Genet. 10:265. doi: 10.3389/fgene.2019.00265

Next-generation sequencing technologies allow to measure somatic mutations in a large number of patients from the same cancer type: one of the main goals in their analysis is the identification of mutations associated with clinical parameters. The identification of such relationships is hindered by extensive genetic heterogeneity in tumors, with different genes mutated in different patients, due, in part, to the fact that genes and mutations act in the context of *pathways*: it is therefore crucial to study mutations in the context of interactions among genes. In this work we study the problem of identifying subnetworks of a large gene-gene interaction network with mutations associated with survival time. We formally define the associated computational problem by using a score for subnetworks based on the log-rank statistical test to compare the survival of two given populations. We propose a novel approach, based on a new algorithm, called Network of Mutations Associated with Survival (NoMAS) to find subnetworks of a large interaction network whose mutations are associated with survival time. NoMAS is based on the color-coding technique, that has been previously employed in other applications to find the highest scoring subnetwork with high probability when the subnetwork score is additive. In our case the score is not additive, so our algorithm cannot identify the optimal solution with the same guarantees associated to additive scores. Nonetheless, we prove that, under a reasonable model for mutations in cancer, NoMAS identifies the optimal solution with high probability. We also design a holdout approach to identify subnetworks significantly associated with survival time. We test NoMAS on simulated and cancer data, comparing it to approaches based on single gene tests and to various greedy approaches. We show that our method does indeed find the optimal solution and performs better than the other approaches. Moreover, on three cancer datasets our method identifies subnetworks with significant association to survival when none of the genes has significant association with survival when considered in isolation.

Keywords: cancer genomics, survival analysis, network analysis, log-rank statistic, holdout approach

1. INTRODUCTION

Recent advances in next-generation sequencing technologies have enabled the collection of sequence information from many genomes and exomes, with many large human and cancer genetic studies measuring mutations in all genes for a large number of patients of a specific disease (Cancer Genome Atlas Research Network, 2013, 2014; Cancer Genome Atlas Network, 2015; Cancer Genome Atlas Research Network et al., 2017; Raphael et al., 2017). One of the main challenges in these studies is the interpretation of such mutations, in particular the identification of mutations that are clinically relevant. For example, in large cancer studies one is interested in finding somatic mutations that are associated with survival and that can be used for prognosis and therapeutic decisions. One of the main obstacles in finding mutations that are clinically relevant is the large number of mutations present in each cancer genome. Recent studies have shown that each cancer genome harbors hundreds or thousands of somatic mutations (Garraway and Lander, 2013), with only a small number (e.g., $\leq$ 10) of *driver* mutations related to the disease, while the vast majority of mutations are *passenger*, random mutations that are accumulated during the process that leads to cancer but not related to the disease (Vogelstein et al., 2013).

In recent years, several computational and statistical methods have been designed to identify driver mutations and distinguish them from passenger mutations, exploiting data from large cancer studies (Raphael et al., 2014). Many of these methods analyze each gene in isolation, and use different single gene scores (e.g., mutation frequency, clustering of mutations, etc.) to identify significant genes (Dees et al., 2012; Lawrence et al., 2013; Tamborero et al., 2013). While useful in finding driver genes, these methods suffer from the extensive *heterogeneity* of mutations in cancer, with different patients showing mutations in different cancer genes (Kandoth et al., 2013). One of the reasons of such mutational heterogeneity is the fact that driver mutations do not target single genes but rather *pathways* (Vogelstein et al., 2013), groups of interacting genes that perform different functions in the cell. Several methods have been recently proposed to identify significant groups of interacting genes in cancer (Vandin et al., 2012b; Hofree et al., 2013; Kim et al., 2015; Leiserson et al., 2015a,b; Shrestha et al., 2017). Many of these methods integrate mutations with interactions from genome-scale interaction networks, without restricting to already known pathways, that would hinder the ability to discover new important groups of genes.

In addition to mutation data, large cancer studies often collect also clinical data, including survival information, regarding the patients. An important feature of survival data is that it often contains *censored* measurements (Kalbfleisch and Prentice, 2002): in many studies a patient may be alive at the end of the study or may leave the study before it ends, therefore only a lower bound to the survival of the patient is known. Survival information is crucial in identifying mutations that have a clinical impact. However, the survival information is commonly used only *after* candidate genes or groups of genes have been identified using other methods, as the ones described above, to evaluate the clinical significance of such genes or groups of genes (Cancer Genome Atlas Research Network, 2011; Hofree et al., 2013). Overall, there is a lack of methods that integrate mutations, interaction information, and survival data to directly identify groups of interacting genes associated with survival.

The field of survival analysis has produced an extensive literature on the analysis of survival data, in particular for the comparison of the survival of two given populations (sets of samples) (Kalbfleisch and Prentice, 2002). The most commonly used test for this purpose is the log-rank test (Mantel, 1966; Peto and Peto, 1972). In genomic studies we are not given two populations, but a single set of samples, and are required to identify mutations that are associated with survival. The log-rank test can be used to this end to identify single genes associated with survival time by comparing the survival of the patients with a mutation in the gene with the survival of the patients with no mutation in the gene. The other commonly used test, the Cox Proportional-Hazards model (Kalbfleisch and Prentice, 2002), is equivalent to the log-rank test when the association of a binary feature with survival is tested, as it is in the case of interest to genomic studies. For a given group of genes, one can *assess* the association of mutations in the genes of the group with survival by comparing the survival of the patients having a mutation in at least one of the genes with the survival of the patients with no mutation in the genes. However, this approach cannot be used to *discover* sets of genes, since one would have to screen all possible subsets of genes and test their association with survival, and the number of subsets of genes to screen is enormous even considering only groups of genes interacting in a protein interaction network (e.g., there are $> 10^{15}$ groups of 8 interacting genes in HINT+HI2012 network; Leiserson et al., 2015b).

In this paper we study the problem of finding sets of interacting genes with mutations associated to survival using data from large cancer sequencing studies and interaction information from a genome-scale interaction network. We focus on the widely used log-rank statistic as a measure of the association between mutations in a group of genes and survival. Our contribution is in five parts: first, we formally define the problem of finding the set of k genes whose mutations show the maximum association to survival time by using the log-rank statistic as a score for a set of genes: we show that such problem is NP-hard. We show that the problem remains hard when the set of k genes is required to form a connected subnetwork in a large graph with at least one node of large degree (*hub*). Second, we propose an efficient algorithm, Network of Mutations Associated with Survival (NoMAS), based on the color-coding technique, to identify subnetworks associated with survival time. Color-coding has been previously used to find high scoring graphs for bioinformatics applications (Dao et al., 2011; Hormozdiari et al., 2015) when the score for a subnetwork is *set additive* (i.e., the score of a subnetwork is the sum of the scores of the genes in the subnetwork). In our case the log-rank statistic is not set additive, and we prove that there is a family of instances for which our algorithm cannot identify the optimal

solution. Nonetheless, we prove that, under a reasonable model for mutations in cancer, our algorithm identifies the optimal solution with high probability. Third, we test our algorithm on simulated data and on data from three large cancer studies from The Cancer Genome Atlas (TCGA). On simulated data, we show that our algorithm does find the optimal solution while being much more efficient than the exhaustive algorithm that screens all sets of genes. On cancer data, we show that our algorithm finds the optimal solution for all values of k for which the use of the exhaustive algorithm is feasible, and identifies better solutions (in terms of association to survival) than a greedy algorithm similar to the one used in Reimand and Bader (2013). Fourth, to strengthen the statistical reliability of NoMAS's results, we employ a holdout scheme, splitting the patients dataset in two parts, a *training* set and a *holdout* set. While solutions of the NoMAS are computed on the former, the assessments of their statistical significance are performed on the latter, thus providing a correction for the multiple hypothesis testing performed on the training set. Finally, we show that NoMAS identifies better solutions than using an (additive) score (i.e., the same gene score used in Vandin et al., 2012a) for a set of genes. For the cancer datasets, we show that our algorithm identifies novel groups of genes associated with survival where none of them is associated with survival when considered in isolation. The work is organized as follows: in section 2 we provide the description of the model and NoMAS; section 3 presents the analysis of the algorithm (section 3.1), including the analysis under a reasonable model for mutations in cancer and analysis of our experiments on both simulated and real data (section 3.2); finally section 4 presents the discussion of our results. Details for our theoretical results are given in **Supplementary Material**.

2. MATERIALS AND METHODS

In this section we present the model we consider, our algorithm NoMAS, and the tests we have designed to assess the statistical significance of the results.

2.1. Computational Problem

In survival analysis, we are given two populations (i.e., sets of samples) P_0 and P_1, and for each sample $i \in P_0 \cup P_1$ we have its survival data: i) the survival time t_i and ii) the censoring information c_i, where $c_i = 1$ if t_i is the exact survival time for sample i (i.e., sample i is not censored), and $c_i = 0$ if t_i is a lower bound to the survival time for sample i (i.e., sample i is censored). Let m_0 be the number of samples in P_0, m_1 be the number of samples in P_1, and $m = m_0 + m_1$ be the total number of samples. Without loss of generality, the samples are $\{1, 2, \ldots, m\}$, the survival times are $t = 1, 2, \ldots, m$, with $t_i = i$ (i.e., the samples are sorted by increasing values of survival), and we assume that there are no ties in survival times. The survival data is represented by two vectors $\mathbf{c}$ and $\mathbf{x}$, with c_i representing the censoring information for sample i, and x_i represents the population information: $x_i = 1$ if sample i is in population P_1, and $x_i = 0$ otherwise. Given the survival data for two populations P_0 and P_1, the significance in the difference of survival between P_0 and P_1 can be assessed by the widely used log-rank test (Mantel, 1966; Peto and Peto, 1972). The log-rank statistic is

$$V(\mathbf{x}, \mathbf{c}) = \sum_{j=1}^{m} c_j \left(x_j - \frac{m_1 - \sum_{i=1}^{j-1} x_i}{m - j + 1} \right) \quad (1)$$

Under the (null) hypothesis of no difference in survival between P_0 and P_1, the log-rank statistic asymptotically follows a normal distribution $\mathcal{N}(0, \sigma^2)$, where the standard deviation[1] is given by: $\sigma(\mathbf{x}, \mathbf{c}) = \sqrt{\frac{m_0 m_1}{m(m-1)} \left(\left(\sum_{j=1}^{m} c_j \right) - \sum_{j=1}^{m} c_i \frac{1}{m-j+1} \right)}$. Thus the normalized log-rank statistic, defined as $\frac{V(\mathbf{x},\mathbf{c})}{\sigma(\mathbf{x},\mathbf{c})}$, asymptotically follows a standard normal $\mathcal{N}(0, 1)$ distribution, and the deviation of $\frac{V(\mathbf{x},\mathbf{c})}{\sigma(\mathbf{x},\mathbf{c})}$ from 0 is a measure of the difference in survival between P_0 and P_1.

In genomic studies, we are given mutation data for a set $\mathcal{G}$ of n genes in a set $\mathcal{P}$ of m samples, represented by a mutation matrix M with $M_{i,j} = 1$ if gene i is mutated in patient j and $M_{i,j} = 0$ otherwise. We are also given survival data (survival time and censoring information) for all the m samples. Given a set $\mathcal{S} \subset \mathcal{G}$ of genes, we can assess the association of mutations in the set $\mathcal{S}$ with survival by comparing the survival of the population $P_1^{\mathcal{S}}$ of samples with a mutation in at least one gene of $\mathcal{S}$ and the survival of the population $P_0^{\mathcal{S}}$ of samples with no mutation in the genes of $\mathcal{S}$. That is, $P_0^{\mathcal{S}} = \{j \in \mathcal{P} : \sum_{i \in \mathcal{S}} M_{i,j} = 0\}$ and $P_1^{\mathcal{S}} = \{j \in \mathcal{P} : \sum_{i \in \mathcal{S}} M_{i,j} > 0\}$.

Given the set $\mathcal{G}$ of all genes for which mutations have been measured, we are interested in finding the set $\mathcal{S} \subset \mathcal{G}$ with $|\mathcal{S}| = k$ that has maximum association with survival by finding the set $\mathcal{S}$ that maximizes the absolute value of the normalized log-rank statistic. Given a set $\mathcal{S}$ of genes, let $\mathbf{x}^{\mathcal{S}}$ be a $0 - 1$ vector, with $x_i^{\mathcal{S}} = 1$ if at least one gene of $\mathcal{S}$ is mutated in patient i, and $x_i^{\mathcal{S}} = 0$ otherwise. The normalized log-rank statistic for the set $\mathcal{S}$ is then $\frac{V(\mathbf{x}^{\mathcal{S}},\mathbf{c})}{\sigma(\mathbf{x}^{\mathcal{S}},\mathbf{c})}$. Note that for a given set of patients the censoring information $\mathbf{c}$ is fixed, therefore we can consider the log-rank statistic as a function $V(\mathbf{x}^{\mathcal{S}})$ of $\mathbf{x}^{\mathcal{S}}$ only. Analogously, we can rewrite $\sigma(\mathbf{x}^{\mathcal{S}}, \mathbf{c}) = \sigma(\mathbf{x}^{\mathcal{S}}) f(\mathbf{c})$, where $\sigma(\mathbf{x}^{\mathcal{S}}) = \sqrt{m_1(m - m_1)}$ with $m_1 = |P_1^{\mathcal{S}}|$, and $f(\mathbf{c}) = \sqrt{\frac{1}{m(m-1)} \left(\left(\sum_{j=1}^{m} c_j \right) - \sum_{j=1}^{m} c_j \frac{1}{m-j+1} \right)}$ does not depend on $\mathbf{x}^{S}$ and is fixed given $\mathbf{c}$.

To identify the set of k genes most associated with survival, we can then consider the score $|w(\mathcal{S})| = \left| \frac{V(\mathbf{x}^{\mathcal{S}})}{\sigma(\mathbf{x}^{\mathcal{S}})} \right|$. For ease of exposition in what follows we consider the score $w(\mathcal{S})$, corresponding to a one tail log-rank test for the identification of sets of genes with mutations associated with reduced survival; the identification of sets of genes with mutations associated with increased survival is done in an

[1] In the literature two different standard deviations (corresponding to two related but different null distributions, permutational and conditional) have been proposed for the normal approximation of the distribution of the log-rank statistic; we have previously shown (Vandin et al., 2015) that the one we use here (corresponding to the permutational distribution) is more appropriate for genomic studies.

analogous way by maximizing the score $-w(\mathcal{S})$. We define the following problem.

The max k-set log-rank problem: Given a set $\mathcal{G}$ of genes, an $n \times m$ mutation matrix M and the survival information (time and censoring) for the m patients in M, find the set $\mathcal{S} \subset \mathcal{G}$ of k genes maximizing $w(\mathcal{S})$.

We have the following.

Theorem 1. *The max k-set log-rank problem is NP-hard.*

We now define the max connected k-set log-rank problem that is analogous to the max k-set log-rank problem but requires feasible solutions to be connected subnetworks of a given graph $\mathcal{I}$, representing gene-gene interactions.

The max connected k-set log-rank problem: Given a set $\mathcal{G}$ of genes, a graph $\mathcal{I} = (\mathcal{G}, E)$ with $E \subseteq \mathcal{G} \times \mathcal{G}$, an $n \times m$ mutation matrix M and the survival information (time and censoring) for the m patients in M, find the set $\mathcal{S}$ of k genes maximizing $w(\mathcal{S})$ with the constraint that the subnetwork induced by $\mathcal{S}$ in $\mathcal{I}$ is connected.

If $\mathcal{I}$ is the complete graph, the max connected k-set log-rank problem is the same as the max k-set log-rank problem. Thus, the max connected k-set log-rank problem is NP-hard for a general graph. However, we can prove that the problem is NP-hard for a much more general class of graphs.

Theorem 2. *The max connected k-set log-rank problem on graphs with at least one node of degree $O\left(n^{\frac{1}{c}}\right)$, where $c > 1$ is constant, is NP-hard.*

2.2. Algorithm NoMAS

We design a new algorithm, Network of Mutations Associated with Survival (NoMAS)[2], to solve the max connected k-set log-rank problem. The algorithm is based on an adaptation of the color-coding technique (Alon et al., 1994). Our algorithm is analogous to other color-coding based algorithms that have been used before to identify subnetworks associated with phenotypes in other applications where the score is additive (Dao et al., 2011; Hormozdiari et al., 2015).

Figure 1 provides an overview of NoMAS. The input to NoMAS is an undirected graph $G = (V, E)$, an $n \times m$ mutation matrix M, and the survival information $\mathbf{x}, \mathbf{c}$ for the m patients in M. NoMAS first identifies a subnetwork $\mathcal{S}$ with high weight $\frac{w(\mathbf{x}^{\mathcal{S}})}{\sigma(\mathbf{x}^{\mathcal{S}})}$. To identify a subnetwork of high weight, the algorithm proceeds in iterations. In each iteration NoMAS colors G with k colors by assigning to each vertex v a color $\mathcal{C}(v) \in \{1, \ldots, k\}$ chosen uniformly at random. For a given coloring of G, a subnetwork $\mathcal{S}$ is said to be *colorful* if all vertices in $\mathcal{S}$ have distinct colors. The *colorset* of $\mathcal{S}$ is the set of colors of the vertices in $\mathcal{S}$. Note that the number of different colorsets (subsets of $\{1, \ldots, k\}$) is 2^k. In each iteration the algorithm efficiently identifies high-scoring colorful subnetworks, and at the end the highest-scoring subnetwork among all iterations is reported.

[2]The implementation of NoMAS is available at https://github.com/VandinLab/NoMAS

Consider a given coloring of G. Let W be a $(2^k - 1) \times |V|$ table with a row for each non-empty colorset and a column for each vertex in G. Entry $W(T, u)$ stores the set of vertices of one connected colorful subnetwork that has colorset T and includes vertex u. Entries of W can be filled by dynamic programming. For colorsets of size 1, the corresponding rows in W are filled out trivially: $W(\{\alpha\}, u) = \{u\}$ if $\alpha = \mathcal{C}(u)$, and $W(\{\alpha\}, u) = \emptyset$ otherwise.

For entry $W(T, u)$ with $|T| \geq 2$, NoMAS computes $W(T, u)$ by combining a previously computed $W(Q, u)$ for u with another previously computed $W(R, v)$ where v is a neighbor of u in G, ensuring that the resulting subnetwork is connected and contains u. Colorfulness is ensured by selecting Q and R such that $Q \cap R = \emptyset$ and $Q \cup R = T$, and in turn ensures that $W(T, u)$ contains $|T|$ distinct vertices. Note that for a given T the choice of Q uniquely defines R. Thus, for each neighbor v of u there are (at most) $2^{|T|-1}$ possible combinations. Let $\mathcal{S}'(T, u)$ be the set of all colorful subnetworks that can be obtained by combining an entry $W(Q, u)$ for u and an appropriate entry $W(R, v)$ for a neighbor v of u so that $Q \cup R = T, Q \cap R = \emptyset$. That is: $\mathcal{S}'(T, u) = \bigcup_{\substack{v:(u,v)\in E \\ Q\cup R=T, Q\cap T=\emptyset}} \left\{W(Q, u) \cup W(R, v)\right\}$ (in the definition of $\mathcal{S}'(T, u)$ we assume that the union with $\emptyset$ returns $\emptyset$). $W(T, u)$ stores the element of $\mathcal{S}'(T, u)$ with largest value of our objective function, that is $W(T, u) = \arg\max_{\mathcal{S}\in\mathcal{S}'(T,u)} w(\mathcal{S})$. At the end, the best solution is identified by finding the entry of W of maximum weight. Analogously, NoMAS identifies sets that minimize $w(\mathcal{S})$ (sets associated to increased survival) by maximizing the score $-w(\mathcal{S})$. (See Appendix for pseudo code and illustrations of the working of NoMAS).

Parallelization

The computation of W is parallelized using $N \leq |V|$ processors. All entries of W are kept in shared memory and $|V|/N$ unique columns uniformly at random are assigned to each processor. Entries of W are computed in order of increasing colorset sizes. We define the i-th *colorset group* as the set of all $\binom{k}{i}$ colorsets of size i. We exploit the fact that the rows within the i-th colorset group are computed by reading entries exclusively from rows belonging to colorset groups $< i$. When a processor has finished the rows of the i-th colorset group it waits for the other processors to do the same. When the last processor completes the i-th colorset group, all N processors can safely begin to compute rows of colorset group $i+1$. In total, k synchronization steps are carried out, one for each colorset group.

2.3. Statistical Significance

We designed two procedures to assess the statistical significance of the results found by NoMAS: the first is based on permutation testing, while the second uses a holdout approach.

Permutation Testing

After identifying the best solution $\mathcal{S}$ for the mutation matrix M, NoMAS can assess its statistical significance by i) estimating the p-value $p(\mathcal{S})$ for the log-rank statistic (using a Monte-Carlo estimate with 10^8 samples), and then ii) using a permutation test in which $\mathcal{S}$ is compared to the best solution $\mathcal{S}^p$ for the

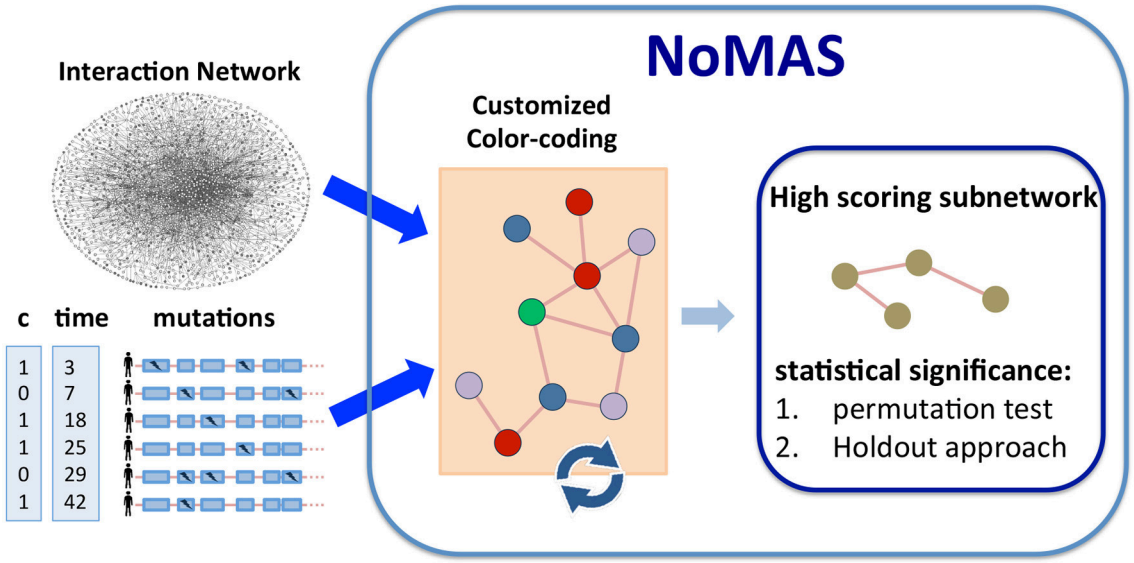


FIGURE 1 | Algorithm NoMAS. Given alteration data and survival information (time and censoring status) for a set of patients, NoMAS employs a color coding approach to identify subnetworks with mutations associated with survival time, i.e., with high log-rank statistic, and then assesses the statistical significance of the subnetworks using (i) permutation testing and (ii) a holdout approach.

mutation matrix M^p obtained by randomly permuting the rows of M. A total of 100 permutations are performed and the *permutation* p-value is recorded as the ratio of permutations in which $w(S^p) \geq w(\mathcal{S})$. While the p-value from the log-rank test reflects the association between mutations in the subnetwork and survival, the permutation p-value assesses whether a subnetwork with association with survival at least as extreme as the one observed in the input data can be observed when the genes are placed randomly in the network. Note that we can identify multiple solutions by considering different entries of W (even if the same solution may appear in multiple entries of W), and we obtain a permutation p-value for the i-th top scoring solution by comparing its score with the score of the i-th top scoring solution in the permuted datasets.

Holdout Method

We designed a holdout method to strengthen the statistical robustness of the results produced by NoMAS. We split the dataset in two parts, called *training* and *holdout*, and then run NoMAS on the former, obtaining subnetworks with high weight. The p-value of these subnetworks is then computed with a Monte-Carlo procedure estimate with 10^8 samples on the holdout dataset. More in detail, assuming that a set P of m patients is analyzed, let v be a parameter with value in $(0, 1)$ that represents the proportion of data in the training set: we partition P into two parts, P_t and P_h, sized $m_t = \lfloor mv \rfloor$ and $m_h = m - m_t$ respectively. In order to preserve the survival distribution in both the training and the holdout set, the partition is performed over each of g temporal intervals of the same length, where g is a parameter to be passed in input by the user. The sets P_t and P_h are obtained by the union of the corresponding sets in each interval. Once we obtain the partition of P into P_t and P_h, NoMAS is executed over the population P_t and p-value of the found solution is computed over P_h.

3. RESULTS

3.1. Analysis of NoMAS

We consider the performance of NoMAS excluding the statistical significance testing. The log-rank statistic $w(\mathcal{S})$ is computed in time $O(m_1) \in O(m)$. The total time complexity for computing a single entry $W(T, u)$ is then bounded by $O\left(m\deg(u)2^{|T|-1}\right) \in O\left(m\deg(u)2^k\right)$, where $deg(u)$ is the degree of u in G. Given a coloring of G, the computation of the entire table can thus be performed in time $O\left(2^k \sum_{u \in V} m\deg(u)2^k\right) \in O\left(m|E|4^k\right)$. If L iterations are performed, then the complexity of the algorithm is $O\left(Lm|E|4^k\right)$.

Let OPT be the optimal solution. If the score $w(\mathcal{S})$ was set additive, as the scores considered in previous applications of color-coding for optimization problems on graphs, to discover OPT it would be sufficient that OPT be colorful, that happens with probability $k!/k^k \geq e^{-k}$ for each random coloring. Therefore $O\left(\ln(1/\delta)e^k\right)$ iterations would be enough to ensure that the probability of OPT not being discovered is $\leq \delta$, resulting in an overall time complexity of $O\left(m\ln(1/\delta)|E|(4e)^k\right)$.

However, our score $w(\mathcal{S})$ is not set additive [e.g., if two genes in $\mathcal{S}$ have a mutation in the same patient the weight of the patient is considered only once in $w(\mathcal{S})$]. Therefore, while OPT being colorful is still a necessary condition for the algorithm to identify OPT, the colorfulness of OPT is not a sufficient condition. In fact, we have the following.

Proposition 1. *For every $k \geq 3$ there is a family of instances of the max connected k-set log-rank problem and colorings for which OPT is not found by NoMAS when it is colorful.*

Even more, we prove that when mutations are placed arbitrarily then for every subnetwork $\mathcal{S}$ and a given coloring of $\mathcal{S}$, *any* color-coding algorithm that adds subnetworks of size k to W by merging neighboring subnetworks of size $< k$ could be "fooled" to not add $\mathcal{S}$ to W by simply adding 3 vertices to G and assigning them a specific color.

Theorem 3. *For any optimal colorful connected subnetwork $\mathcal{S}$ of size $k \geq 3$ and any color-coding algorithm $\mathcal{A}$ which obtains subnetworks with colorsets of cardinality i by combining 2 subnetworks with colorsets of cardinality $< i$, by adding 3 neighbors to $\mathcal{S}$ we have that $\mathcal{A}$ may not discover S.*

Intuitively, Proposition 1 and Theorem 3 show that if mutations are placed adversarially (and the optimal solution *OPT* has many neighbors), our algorithm may not identify *OPT*. However, we prove that our algorithm identifies the optimal solution under a generative model for mutations, that we deem the *Planted Subnetwork Model*. We consider $w(\mathcal{S})$ as the unnormalized version of the log-rank statistic. In this model: i) there is a subnetwork $\mathcal{D}$, $|\mathcal{D}| = k$, with $w(\mathcal{D}) \geq cm$, for a constant $c > 0$; ii) each gene $g \in \mathcal{D}$ is such that $w(\mathcal{D}) - w(\mathcal{D} \setminus \{g\}) \geq \frac{c'm}{k}$, for a constant $c' > 0$; iii) for each gene $g \in \mathcal{D}$: $w(\{g\}) > 0$; iv) for each gene $\hat{g} \notin \mathcal{D}$, $\hat{g}$ is mutated with probability p_g in each patient, independently of all other events (and of survival time and censoring status).

Intuitively: (3.1) above states that the subnetwork $\mathcal{D}$ has mutations associated with survival; (3.1) states that each gene $g \in \mathcal{D}$ contributes to the association of mutations in $\mathcal{D}$ to survival; (3.1) states that each gene $g \in \mathcal{D}$ should have the same association to survival (increased or decreased) as $\mathcal{D}$; and (3.1) states that all mutations outside $\mathcal{D}$ are independent of all other events (including survival time and censoring of patients).

We show that when enough samples are generated from the model above, our algorithm identifies the optimal solution with the same probability guarantee given by the color-coding technique for additive scores.

Theorem 4. *Let M be a mutation matrix corresponding to m samples from the Planted Subnetwork Model. If $m \in \Omega\left(k^4(k+\varepsilon)\ln n\right)$ for a given constant $\varepsilon > 0$ and $O\left(\ln(1/\delta)e^k\right)$ color-coding iterations are performed, then our algorithm identifies the optimal solution $\mathcal{D}$ to the max connected k-set log-rank with probability $\geq 1 - \frac{1}{n^\varepsilon} - \delta$.*

3.2. Experimental Results

We assessed the performance of NoMAS by using simulated and cancer data. We compared NoMAS to the exhaustive algorithm that identifies the subnetwork of k vertices with the highest score $w(S)$ for the values of k for which we could run the exhaustive algorithm (we implemented a parallelized version of the algorithm described in Maxwell et al., 2014 to efficiently enumerate all connected subnetworks), to three variants of a greedy algorithm similar to the one from Reimand and Bader (2013), and to the use of a score given by the sum of single gene scores. Cancer data is obtained from The Cancer Genome Atlas (TCGA). In particular, we consider somatic mutations (single nucleotide variants and small indels) for 268 samples of glioblastoma multiforme (GBM), 315 samples of ovarian adenocarcinoma (OV) and 174 samples of lung squamous cell carcinoma (LUSC) for which survival data is available.

For all our experiments we used as interaction graph G the graph derived from the application of a diffusion process on the HINT+HI2012 network[3], a combination of the HINT network (Das and Yu, 2012) and the HI-2012 (Yu et al., 2011) set of protein-protein interactions, previously used in Leiserson et al. (2015a). The details of the diffusion process are described in Leiserson et al. (2015a). In brief, for two genes g_i, g_j the diffusion process gives the amount of heat $h(g_i, g_j)$ observed on g_j when g_i has one mutation, and the amount of heat $h(g_j, g_i)$ observed on g_i when g_j has one mutation. The graph used for our analyses is obtained retaining an edge between g_i and g_j if $\max\{h(g_i, g_j), h(g_j, g_i)\} \geq 0.012$. The resulting graph has 9,859 vertices and 42,480 edges, with the maximum degree of a node being 438. In all our experiments we removed mutations in genes mutated in < 3 of the samples. For cancer data, this resulted in 890 mutated genes removed in GBM, 780 in OV, and 2,915 in LUSC. The machine, on which all our experiments were carried out, consists of two CPUs of the type Intel Xeon E5-2698 v3 (2.30 GHz), each with 16 physical cores, for a total of 64 virtual cores, and 16 banks of 32 GB DDR4 (2,133 MHz) memory modules for a total of 512 GB of memory.

The remaining of the section is organized as follow: section 3.2.1 presents the results on simulated data, while section 3.2.2 presents the results on cancer data.

3.2.1. Simulated Data

We assess the performance of NoMAS on simulated data generated under the Planted subnetwork Model. The subnetwork $\mathcal{D} \subset \mathcal{G}, |\mathcal{D}| = k$ associated with survival is generated by a random walk on the graph G. We model the association of $\mathcal{D}$ to survival by mutating with probability p one gene of $\mathcal{D}$ chosen uniformly at random in each sample among the $\frac{m}{4}$ of lowest survival. All other genes in $\mathcal{D}$ are mutated independently with probability 0.01 in all samples, to simulate passenger mutations (not associated with survival) in $\mathcal{D}$ (Lawrence et al., 2013). For genes in $\mathcal{G} \setminus \mathcal{D}$, we used the same mutation frequencies observed in the GBM study, and mutate each gene independently of all other events.

We fixed $k = 5$ and considered the values of $p \in \{0.5, 0.75, 0.85\}$ and $m \in \{268, 500, 750, 1,000\}$. We kept the same ratio of censored observations as in GBM and chose the censored samples uniformly among all samples. For every pair (p, m) we performed 100 simulations, running NoMAS on the dataset with $L = 256$ color-coding iterations, and recorded whether NoMAS reported $\mathcal{D}$ as the highest scoring subnetwork. Results are shown in **Figure 2A**. For sample sizes similar to the currently available ones, NoMAS frequently reports $\mathcal{D}$ as the highest scoring solutions when there is a quite strong association

[3]http://compbio-research.cs.brown.edu/pancancer/hotnet2/

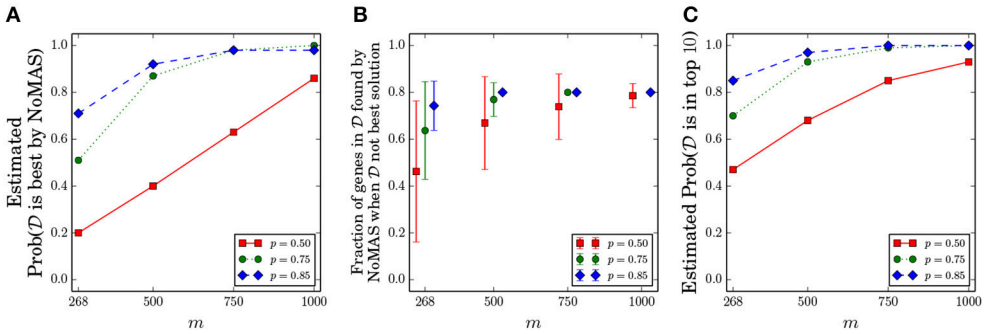


FIGURE 2 | Results of NoMAS on simulated data from the Planted Subnetwork Model. One hundred datasets were generated for each pair (m,p), where m is the number of samples and for different probabilities p of mutations in the set $\mathcal{D}$ of genes associated with survival. **(A)** Probability that $\mathcal{D}$ is reported as the highest scoring solution by NoMAS. **(B)** Ratio of genes from the set $\mathcal{D}$ that are in the best solution when $\mathcal{D}$ is not the highest scoring solution by NoMAS. **(C)** Probability that $\mathcal{D}$ is among the top-10 solutions reported by NoMAS. All probabilities are estimated from the simulated datasets.

of $\mathcal{D}$ with survival ($p \geq 0.85$), but for $m = 1,000$ the highest scoring subnetwork reported by NoMAS is $\mathcal{D}$ in $> 80\%$ of the cases even for $p = 0.5$. **Figure 2B** shows that even when NoMAS does not report $\mathcal{D}$ as the highest scoring solution, the solution reported by NoMAS contains mostly genes that are in $\mathcal{D}$, even for current sample size (e.g., on average 74% of the genes in the $\mathcal{D}$ are reported by NoMAS for $m = 268$ and $p = 0.85$ even when $\mathcal{D}$ is not the highest scoring solution by NoMAS). Finally, we assessed whether $\mathcal{D}$ would be among the highest scoring solutions in the table W computed by NoMAS: **Figure 2C** shows that by considering the top-10 solutions W the chances to identify $\mathcal{D}$ increase substantially even for $m = 268$ and $p = 0.5$, with most configurations having > 0.8 probability of finding $\mathcal{D}$ in the top-10 solutions by NoMAS. For a fixed $p = 0.75$ and for each value of m we assessed whether NoMAS identified the optimal solution even when it was not $\mathcal{D}$ (an event not excluded in the Planted subnetwork Model) and found that for $m \geq 500$ NoMAS reported the optimal solution in 10 out of 10 cases (for $m = 268$ NoMAS identified the optimal solution 9 out of 10 times). These results show that NoMAS does indeed find the optimal solution in almost all cases even for sample sizes currently available (while the theoretical analysis of section 3.1 suggests that much larger sample sizes are required) and it can be used to identify $\mathcal{D}$ or the majority of it by considering the top-10 highest scoring solutions.

3.2.2. Cancer Data

We assessed the performance of NoMAS on the GBM, OV, and LUSC datasets. We first assessed whether NoMAS identified the optimal solution by comparing the highest scoring solution reported by NoMAS with the one identified by using the exhaustive algorithm for $k = 2, 3, 4, 5$. In all cases we found that NoMAS does identify the optimal solution, while requiring much less running time compared to the exhaustive algorithm (**Supplementary Figure 2**). For $k > 5$ we could not run the exhaustive algorithm, while the runtime of NoMAS is still reasonable. The runtime of NoMAS can be greatly improved by using the parallelization strategy described in section 2.2 (**Supplementary Figure 3**). We therefore used NoMAS to find subnetworks of size $k = 6$ and $k = 8$. We also considered two modifications of NoMAS that solve some easy cases where NoMAS may not identify the highest scoring solution due to its subnetwork merging strategy (see Appendix for a description and pseudocode of the modifications). We run both modifications on GBM, OV, and LUSC for $k = 6, 8$ (using the same colorings used by the original version of NoMAS): in all cases the modified versions of NoMAS did not report subnetworks with higher scores than the ones from the original version of NoMAS. We also note that the original version of NoMAS is significantly faster in practice than its two modifications (**Supplementary Figure 3**) and, therefore, we used the original version of NoMAS in the remaining experiments.

We also compared NoMAS with three different greedy strategies for the max connected k-set log-rank problem. All three algorithms build solutions starting from each node $u \in G$ and, in iterations, by adding nodes to the current solution $\mathcal{S}$, diversifying in the way they enlarge the current subnetwork $\mathcal{S}$ of size $1 \leq i < k$. (See Appendix for a description of the three greedy strategies). We run the three greedy algorithms on GBM, OV, and LUSC for $k = 4, 5, 6, 8$. For each dataset we compared the resulting subnetworks with the ones identified by NoMAS. Results are shown in **Figure 3**. In almost all cases we found that NoMAS discovered subnetworks with higher score than the subnetworks found by using greedy strategies, even if in some cases there is a greedy strategy that identifies the same subnetworks for all values of k. The difference in score increases as k increases, showing the ability of NoMAS to discover better solutions for larger values of k, with the main expense being the running time of NoMAS as opposed to the greedy strategies (**Supplementary Figure 4**). We also assessed whether the fact that greedy strategies discover lower scoring solutions than NoMAS has an impact on the estimate of the p-value in the permutational test. We considered the top-10 scoring solutions (corresponding to 10 different starting nodes $u \in G$) discovered by the best greedy strategy in the GBM dataset and computed the permutational p-value for each solution by generating 100 permuted datasets either using the (same) greedy strategy or NoMAS for (with only 32 iterations on the permuted data). **Supplementary Figure 1**

shows a comparison of the distribution of the p-values. As we can see, the greedy strategy incorrectly underestimate the permutational p-values for the solutions, due to the greedy algorithm not being able to identify solutions of score as high as NoMAS in the permuted datasets. The use of the greedy algorithms would then lead to both 1. identify solutions in real data with lower association to survival compared to NoMAS and 2. wrongly estimate their permutational p-value as more significant than it is.

Finally, we compared NoMAS with the use of an (additive) score that sums single gene scores (similar to the ones used in Vandin et al. (2012a). For each gene $g \in G$ we computed the p-value $p(g)$ for the association of g with survival using the log-rank test and defined $a(\mathcal{S}) = \sum_{g \in \mathcal{S}} -\log_{10} p(g)$. We then partitioned the genes according to their association with increased survival or with decreased survival and modified our algorithm to look for high scoring solutions in a partition using score $a(\mathcal{S})$. Results are in **Figure 3**. We found that NoMAS outperforms the use of a

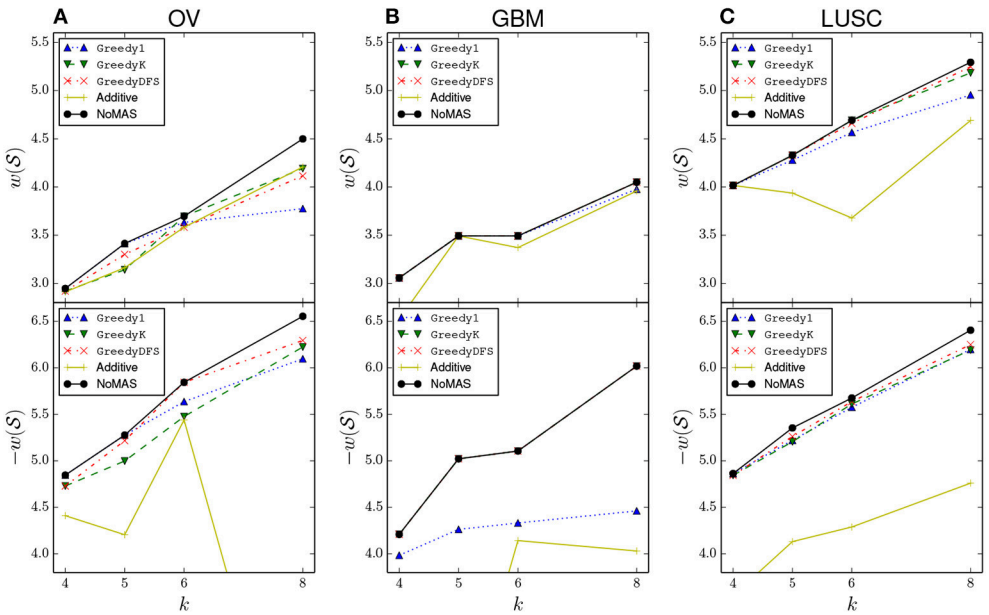


FIGURE 3 | Comparison of the normalized log-rank statistic of the best solution reported by NoMAS, by greedy algorithms (see Appendix for the description), and by the algorithm that uses an additive scoring function $a(\mathcal{S})$ (denote by "additive" in the plots). To maintain readability we omit values above -4.0 when considering mutations associated with increased survival. For each datasets the results for the maximization of $w(\mathcal{S})$ (top panel) and the maximization of $-w(\mathcal{S})$ (bottom panel) are shown separately. **(A)** Results for GBM dataset. **(B)** Results for OV dataset. **(C)** Results for LUSC dataset.

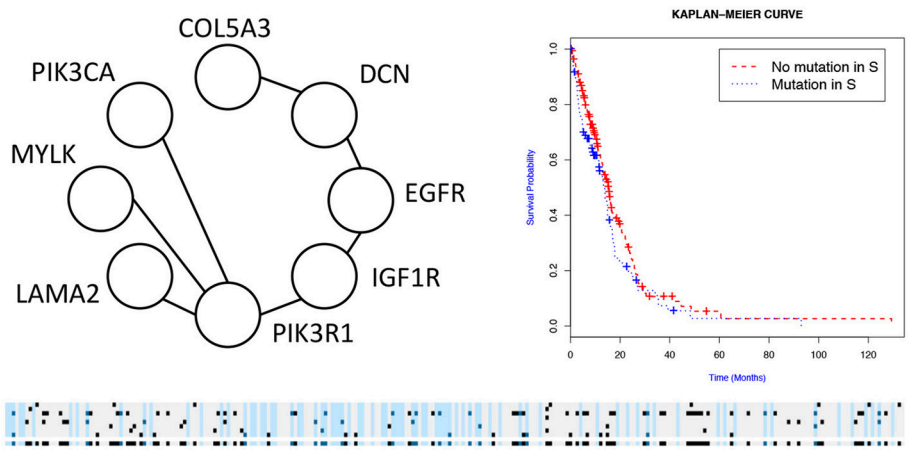


FIGURE 4 | Subnetworks identified by NoMAS on GBM data. Subnetwork S associated with survival in GBM, Kaplan–Meier plot for the samples with mutations in S vs. samples with no mutation in S. The bottom panel shows the mutations in patients for the genes and the entire subnetwork (last row); patients with censored survival are in gray, other patients are in light blue; mutations in patients are show in dark color.

single gene score, with a very large difference for certain values of parameters.

We then used the holdout approach to identify significant subnetworks for GBM, LUSC, and OV, considering the top-10 highest scoring subnetworks found in the training set and compute their p-value in the holdout set. We test all datasets using $k = 3, 4, 8, 256$ iterations of the color coding algorithm. As before, as pre-processing, genes mutated in < 3 samples were eliminated. NoMAS identified several subnetworks with significant association to survival. In GBM, for $k = 8$, NoMAS found the subnetwork including COL5A3, DCN, EGFR, IGF1R, LAMA2, MYLK, PIK3R1, and PIK3CA ($p \leq 0.05$; **Figure 4**). None of the genes is associated with survival when considered in isolation. DCN, EGFR, IGF1R, PIK3R1 recur in various metabolic functions related to lipids and enzymes signaling and reception. These genes, together with PIK3CA, MYLK, and LAMA2, are involved in formation and maintenance of biological tissues, in cell movement and migration and cell protection organization. Moreover, EGFR, PIK3R1, and PIK3CA are well-known cancer genes. EGFR, IGF1R, LAMA2, MYLK, PIK3CA, PIK3R1, and MYLK are members of the focal adhesion pathway, whose dynamics are highly altered in cancer cells. In LUSC, NoMAS found the subnetwork including MAD1L1, USP15, and ZNF434 ($p \leq 0.03$; **Figure 5**). None of the genes is associated with survival when considered in isolation. USP15 stabilizes MDM2, a well-known cancer gene, to regulate cancer-cell survival and mediates antitumor T cell responses (Zou et al., 2014), while increased expression of MAD1L1 is associated with poor prognosis in breast cancer (Sun et al., 2013). In OV, NoMAS identified the subnetwork including EP300, NCOA3, NOTCH1, and NOTCH4 ($p \leq 0.1$; **Figure 6**). None of the genes is associated with survival when considered in isolation. These genes are part of a pathway related to RNA metabolic processes and have a role in regulation of epidermis development and cell differentiation within its layers. All genes are also linked to the thyroid hormone signaling pathway, that is related to cell death and DNA damage in ovarian cancer (Shinderman-Maman et al., 2017).

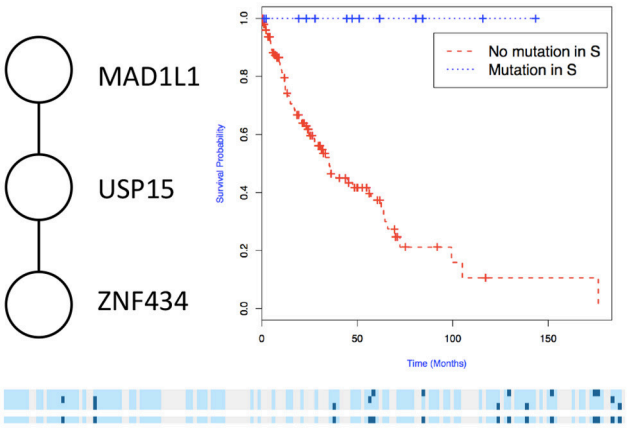


FIGURE 5 | Subnetworks identified by NoMAS on LUSC data. Subnetwork S associated with survival in LUSC, Kaplan–Meier plot for the samples with mutations in S vs. samples with no mutation in S. The bottom panel shows the mutations in patients for the genes and the entire subnetwork (last row); patients with censored survival are in gray, other patients are in light blue; mutations in patients are show in dark color.

4. DISCUSSION

In this work, we study the problem of identifying subnetworks of a large gene-gene interaction network that are associated with survival using mutations from large cancer genomic studies. Few methods have been proposed to identify groups of genes with mutations associated with survival in genomic studies. The work of Vandin et al. (2012a) combines mutations and survival data with interaction information using a diffusion process on graphs starting from gene scores derived from p-values of individual genes, but did not consider the problem of directly identifying groups of genes associated with survival. The work of Reimand and Bader (2013) combines mutation information and patient survival to identify subnetworks of a kinase-substrate interaction network associated with survival. It only focuses on phosphorylation-associated mutations, and the approach is based on a local search algorithm that builds a subnetwork by starting from one seed vertex and then greedily adding neighbors (at distance at most 2) from the seed, extending the approach used in different types of network analyses (Chuang et al., 2007). A similar greedy approach is used by Wu and Stein (2012) to identify groups of genes significantly associated with survival in cancer from gene expression data. For gene expression studies, Chowdhury et al. (2011) proposes an approach to enumerate dysregulated subnetworks in cancer based on an efficient search space pruning strategy, inspired by previous work on the identification of association rules in databases (Smyth and Goodman, 1992). Patel et al. (2013) uses the general approach described in Chowdhury et al. (2011) to identify subnetworks of genes with expression status associated to survival.

Color-coding is a probabilistic method that was originally described for finding simple paths, cycles and other small subnetworks of size k within a given network (Alon et al., 1994). The core of the color-coding technique is the assignment of random colors to the vertices, as a result of which the search space can be reduced, by restricting the subnetworks under consideration to *colorful* ones, those in which each vertex has a distinct color. For the identification of colorful subnetworks, dynamic programming is employed. The process is repeated until the desired subnetwork has been identified, that is having been colorful at least once, with high probability. When the dynamic programming algorithm is polynomial in n and the subnetworks being screened are of size $k \in O(\log n)$, the overall running time of the color-coding method too remains polynomial in n. Color-coding has been previously used to count or search for subgraphs of large interaction networks (Alon et al., 2008; Bruckner et al., 2010). Color-coding has also been used to identify groups of interacting genes in an interaction network that are associated with a phenotype of interest, but restricted to additive scores for sets of genes (i.e., the score of a set is the sum of the scores of the single genes); for example, Dao

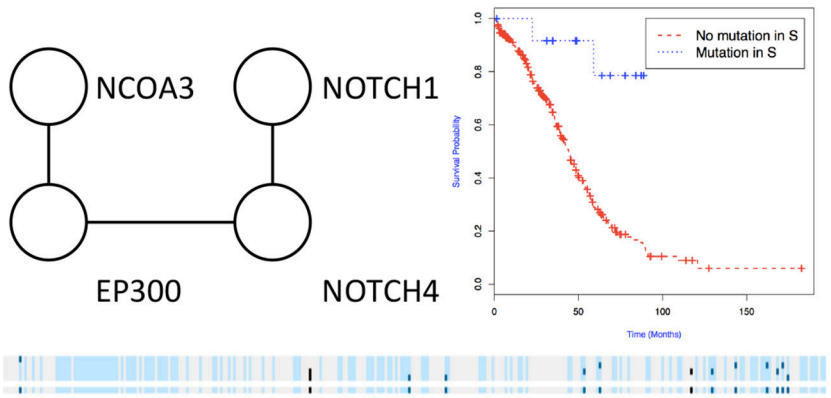


FIGURE 6 | Subnetworks identified by NoMAS on OV data. Subnetwork *S* associated with survival in OV, Kaplan-Meier plot for the samples with mutations in *S* vs. samples with no mutation in *S*. The bottom panel shows the mutations in patients for the genes and the entire subnetwork (last row); patients with censored survival are in gray, other patients are in light blue; mutations in patients are show in dark color.

et al. (2011) uses color-coding to find optimally discriminative subnetwork markers that predict response to chemotherapy from a large interaction network by defining a single gene score as $-\log_{10} d(g)$, where $d(g)$ is the discriminative score for gene g (i.e., a measure of the ability of g to discriminate two classes of patients); similarly, Hormozdiari et al. (2015) uses color-coding to find groups of interacting genes with discriminative mutations in case-control studies, using as gene score the $-\log_{10}$ of the p-value from the binomial test of recurrence of mutations in the cases (while limiting the number of mutations in the controls).

In this work we formally define the associated computational problem, that we call the max connected k-set log-rank problem, by using as score for a subnetwork the test statistic of the log-rank test, one of the most widely used statistical tests to assess the significance in the difference in survival among two populations. We prove that the max connected k-set log-rank problem is NP-hard in general, and is NP-hard even when restricted to graphs with at least one node of large degree. We develop a new algorithm, NoMAS, based on the color-coding technique, to efficiently identify high-scoring subnetworks associated with survival. We prove that even if our algorithm is not guaranteed to identify the optimal solution with the probability given by the color-coding technique (due the non-additivity of our scoring function), it does identify the optimal solution with the same guarantees given by the color-coding technique when the data comes from a reasonable model for mutations and independently of the survival data. Using simulated data, we show that NoMAS is more efficient than the exhaustive algorithm while still identifying the optimal solution, and that our algorithm will identify subnetworks associated with survival when sample sizes larger than most currently available ones, but still reasonable, are available.

We use cancer data from three cancer studies from TCGA to compare NoMAS to approaches based on single gene scores and to greedy methods similar to ones proposed in the literature for the identification of subnetworks associated with survival and for other problems on graphs. Our results show that NoMAS identifies subnetworks with stronger association to survival compared to other approaches, and allows the correct estimation of p-values using a permutation test. Moreover, in two datasets NoMAS identifies two subnetworks associated with survival containing genes previously reported to be important for prognosis in the same cancer type as well as novel genes, while no gene is significantly associated with survival when considered in isolation.

There are many directions in which this work can be extended. First, we only considered single nucleotide variants and indels in our analysis; we plan to extend our method to consider more complex variants (e.g., copy number aberrations and differential methylation) in the analysis. Second, we believe that our algorithm and its analysis could be extended to the identification of subnetworks associated with clinical parameters other than survival time and to case-control studies, but substantial modifications to the algorithm and to its analysis will be required. Third, this work considers the log-rank statistic as a measure of association with survival; another popular test in survival analysis is the use of Cox's regression model (Kalbfleisch and Prentice, 2002). The two tests are identical in the case of two populations, therefore our algorithm identifies subnetworks with high score w.r.t. Cox's regression model as well. However, Cox's regression model allows for the correction for covariates (e.g., gender, age, etc.) in the analysis of survival data. A similar approach could be obtained by stratifying the patients in the log-rank test, but how to efficiently identify subnetworks, and in general combinations of genomic features, associated with survival while correcting for covariates remains a challenging open problem. Fourth, genomic regions other than genes (e.g., regulatory regions) or even other regulatory elements (e.g., microRNAs regulating the expression of driver genes) may be important for survival: the incorporation in our method of alterations in such regions and elements is an interesting direction for future research. Finally, in some studies

the information regarding tumor (sub)clones and their mutations may be available: how to properly integrate such information in our analyses is a challenging direction for further investigation.

AUTHOR CONTRIBUTIONS

FV conceived and designed the study. FA, TH, and FV designed the algorithms, performed the analyses, and wrote the manuscript. FA and TH wrote the software.

FUNDING

This work is supported, in part, by the University of Padova under projects CPDA121378/12, SID 2017, and STARS: Algorithms for Inferential Data mining, and by NSF grant IIS-1247581. The results presented in this manuscript are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This paper was selected for oral presentation at RECOMB 2016 and an abstract is published in the conference proceedings.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00265/full#supplementary-material

REFERENCES

Alon, N., Dao, P., Hajirasouliha, I., Hormozdiari, F., and Sahinalp, S. C. (2008). Biomolecular network motif counting and discovery by color coding. *Bioinformatics* 24, i241–i249. doi: 10.1093/bioinformatics/btn163

Alon, N., Yuster, R., and Zwick, U. (1994). "Color-coding: a new method for finding simple paths, cycles and other small subgraphs within large graphs," in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing* (New York, NY: ACM), 326–335. doi: 10.1145/195058.195179

Bruckner, S., Hüffner, F., Karp, R. M., Shamir, R., and Sharan, R. (2010). Topology-free querying of protein interaction networks. *J. Comput. Biol.* 17, 237–252. doi: 10.1089/cmb.2009.0170

Cancer Genome Atlas Network (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576–582. doi: 10.1038/nature14129

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49. doi: 10.1038/nature12222

Cancer Genome Atlas Research Network (2014). Integrated genomic characterization of papillary thyroid carcinoma. *Cell* 159, 676–690. doi: 10.1016/j.cell.2014.09.050

Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al. (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature* 541:169. doi: 10.1038/nature20805

Chowdhury, S. A., Nibbe, R. K., Chance, M. R., and Koyutürk, M. (2011). Subnetwork state functions define dysregulated subnetworks in cancer. *J. Comput. Biol.* 18, 263–281. doi: 10.1089/cmb.2010.0269

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 27, i205–i213. doi: 10.1093/bioinformatics/btr245

Das, J., and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* 6:92. doi: 10.1186/1752-0509-6-92

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). Music: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Garraway, L. A., and Lander, E. S. (2013). Lessons from the cancer genome. *Cell* 153, 17–37. doi: 10.1016/j.cell.2013.03.002

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E. E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Res.* 25, 142–154. doi: 10.1101/gr.178855.114

Kalbfleisch, J., and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2 Edn*. Hoboken, NJ: Wiley-Interscience.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Kim, Y.-A., Cho, D.-Y., Dao, P., and Przytycka, T. M. (2015). Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 31, i284–i292. doi: 10.1093/bioinformatics/btv247

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015a). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Leiserson, M. D. M., Wu, H.-T., Vandin, F., and Raphael, B. J. (2015b). Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 16:160. doi: 10.1186/s13059-015-0700-7

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* 50, 163–170.

Maxwell, S., Chance, M. R., and Koyutürk, M. (2014). "Efficiently enumerating all connected induced subgraphs of a large molecular network," in *Algorithms for Computational Biology*, eds A.-H. Dediu, C. Martín-Vide, and B. Truthe (Terragona: Springer), 171–182. doi: 10.1007/978-3-319-07953-0_14

Patel, V. N., Gokulrangan, G., Chowdhury, S. A., Chen, Y., Sloan, A. E., Koyutürk, M., et al. (2013). Network signatures of survival in glioblastoma multiforme. *PLoS Comput. Biol.* 9:e1003237. doi: 10.1371/journal.pcbi.1003237

Peto, R., and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc. A* 135, 185–206. doi: 10.2307/2344317

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* 6:5. doi: 10.1186/gm524

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32, 185–203. doi: 10.1016/j.ccell.2017.07.007

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9:637. doi: 10.1038/msb.2012.68

Shinderman-Maman, E., Cohen, K., Moskovich, D., Hercbergs, A., Werner, H., Davis, P. J., et al. (2017). Thyroid hormones derivatives reduce proliferation

and induce cell death and dna damage in ovarian cancer. *Sci. Rep.* 7:16475. doi: 10.1038/s41598-017-16593-x

Shrestha, R., Hodzic, E., Sauerwald, T., Dao, P., Wang, K., Yeung, J., et al. (2017). Hit'ndrive: patient-specific multidriver gene prioritization for precision oncology. *Genome Res.* 27, 1573–1588. doi: 10.1101/gr.221218.117

Smyth, P., and Goodman, R. M. (1992). An information theoretic approach to rule induction from databases. *IEEE Trans. Knowledge Data Eng.* 4, 301–316. doi: 10.1109/69.149926

Sun, Q., Zhang, X., Liu, T., Liu, X., Geng, J., He, X., et al. (2013). Increased expression of mitotic arrest deficient-like 1 (mad1l1) is associated with poor prognosis and insensitive to taxol treatment in breast cancer. *Breast Cancer Res. Treat.* 140, 323–330. doi: 10.1007/s10549-013-2633-8

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* 3:2650. doi: 10.1038/srep02650

Vandin, F., Clay, P., Upfal, E., and Raphael, B. J. (2012a). Discovery of mutated subnetworks associated with clinical data in cancer. *Pac. Symp. Biocomput* 2012, 55–66. doi: 10.1142/9789814366496_0006

Vandin, F., Papoutsaki, A., Raphael, B. J., and Upfal, E. (2015). Accurate computation of survival statistics in genome-wide studies. *PLoS Comput. Biol.* 11:e1004071. doi: 10.1371/journal.pcbi.1004071

Vandin, F., Upfal, E., and Raphael, B. J. (2012b). *De novo* discovery of mutated driver pathways in cancer. *Genome Res.* 22, 375–385. doi: 10.1101/gr.120477.111

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Wu, G., and Stein, L. (2012). A network module-based method for identifying cancer prognostic signatures. *Genome Biol.* 13:R112. doi: 10.1186/gb-2012-13-12-r112

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., et al. (2011). Next-generation sequencing to generate interactome datasets. *Nat. Methods* 8, 478–480. doi: 10.1038/nmeth.1597

Zou, Q., Jin, J., Hu, H., Li, H. S., Romano, S., Xiao, Y., et al. (2014). Usp15 stabilizes mdm2 to mediate cancer-cell survival and inhibit antitumor T cell responses. *Nat. Immunol.* 15:562. doi: 10.1038/ni.2885

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.

Copyright © 2019 Altieri, Hansen and Vandin. 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) and the copyright owner(s) 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 Genetics

METHODS
published: 25 September 2019
doi: 10.3389/fgene.2019.00858

pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks

Ege Ulgen[1], Ozan Ozisik[2] and Osman Ugur Sezerman[1]*

[1] Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey, [2] Department of Computer Engineering, Electrical & Electronics Faculty, Yildiz Technical University, Istanbul, Turkey

Pathway analysis is often the first choice for studying the mechanisms underlying a phenotype. However, conventional methods for pathway analysis do not take into account complex protein-protein interaction information, resulting in incomplete conclusions. Previously, numerous approaches that utilize protein-protein interaction information to enhance pathway analysis yielded superior results compared to conventional methods. Hereby, we present pathfindR, another approach exploiting protein-protein interaction information and the first R package for active-subnetwork-oriented pathway enrichment analyses for class comparison omics experiments. Using the list of genes obtained from an omics experiment comparing two groups of samples, pathfindR identifies active subnetworks in a protein-protein interaction network. It then performs pathway enrichment analyses on these identified subnetworks. To further reduce the complexity, it provides functionality for clustering the resulting pathways. Moreover, through a scoring function, the overall activity of each pathway in each sample can be estimated. We illustrate the capabilities of our pathway analysis method on three gene expression datasets and compare our results with those obtained from three popular pathway analysis tools. The results demonstrate that literature-supported disease-related pathways ranked higher in our approach compared to the others. Moreover, pathfindR identified additional pathways relevant to the conditions that were not identified by other tools, including pathways named after the conditions.

Keywords: pathway analysis, enrichment, tool, active subnetworks, biological interaction network

OPEN ACCESS

Edited by:
Marco Antoniotti,
University of Milano-Bicocca, Italy

Reviewed by:
Ivan Merelli,
Italian National Research Council, Italy
Arnaud Ceol,
Istituto Europeo di Oncologia s.r.l., Italy

****Correspondence:***
Ege Ulgen
egeulgen@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 September 2018*
Accepted: *16 August 2019*
Published: *25 September 2019*

Citation:
Ulgen E, Ozisik O and Sezerman OU (2019) pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. Front. Genet. 10:858. doi: 10.3389/fgene.2019.00858

INTRODUCTION

High-throughput technologies revolutionized biomedical research by enabling comprehensive characterization of biological systems. One of the most common use cases of these technologies is to perform experiments comparing two groups of samples (typically disease versus control) and identify a list of altered genes. However, this list alone often falls short of providing mechanistic insights into the underlying biology of the disease being studied (Khatri et al., 2012). Therefore, researchers face a challenge posed by high-throughput experiments: extracting relevant information that allows them to understand the underlying mechanisms from a long list of genes.

One approach that reduces the complexity of analysis while simultaneously providing great explanatory power is identifying groups of genes that function in the same pathways, i.e., pathway analysis. Pathway analysis has been successfully and repeatedly applied to gene expression (Werner, 2008; Emmert-Streib and Glazko, 2011), proteomics (Wu et al., 2014), and DNA methylation data (Wang et al., 2017).

Most commonly used pathway analysis methods are overrepresentation analysis (ORA) and functional class scoring (FCS). For each pathway, ORA statistically evaluates the proportion of altered genes among the pathway genes against the proportion among a set of background genes. In FCS, a gene-level statistic is calculated using the measurements from the experiment. These gene-level statistics are then aggregated into a pathway-level statistic for each pathway. Finally, the significance of each pathway-level statistic is assessed, and significant pathways are determined.

While they are widely used, there are drawbacks to conventional pathway analysis methods. The statistics used by ORA approaches usually consider the number of genes in a list alone. ORA methods are also independent of the values associated with these genes, such as fold changes or p values. Most importantly, both ORA and FCS methods lack in incorporating interaction information. We propose that directly performing pathway analysis on a gene set is not completely informative because this approach reduces gene-phenotype association evidence by ignoring information on interactions of genes.

We propose a pathway analysis method, which we named pathfindR, that first identifies active subnetworks and then performs enrichment analysis using the identified active subnetworks. For a given list of significantly altered genes, an active subnetwork is defined as a group of interconnected genes in a protein-protein interaction network (PIN) that predominantly consists of significantly altered genes. In other words, active subnetworks define distinct disease-associated sets of interacting genes.

The idea of utilizing PIN information to enhance pathway enrichment results was sought and successfully implemented in numerous studies. Gene Network Enrichment Analysis (GNEA) (Liu et al., 2007) analyzes gene expression data. The mRNA expression of every gene is mapped onto a PIN, and a significantly transcriptionally affected subnetwork is identified via jActiveModules (Ideker et al., 2002). To determine the gene set enrichment, each gene set is then tested for overrepresentation in the subnetwork. In EnrichNet (Glaab et al., 2012), input genes and pathway genes are mapped on a PIN. Using the random walk with restart (RWR) algorithm, distances between input genes and pathway genes are calculated. Enrichment results are obtained by comparing these distances to a background model. In both NetPEA and NetPEA′ (Liu et al., 2017a), initially, the RWR algorithm is used to measure distances between pathways and input gene sets. The significances of pathways are then calculated by comparing against a background model created with two different approaches: a) randomizing input genes (NetPEA) and b) randomizing input genes and the PIN (NetPEA′).

With pathfindR, our aim was likewise to exploit interaction information to extract the most relevant pathways. We aimed to combine together active subnetwork search and pathway enrichment analysis. By implementing this original active-subnetwork-oriented pathway analysis approach as an R package, our intention was to provide the research community with a set of utilities (in addition to pathway analysis, clustering of pathways, scoring of pathways, and visualization utilities) that will be effective, beneficial, and straightforward to utilize for pathway enrichment analysis exploiting interaction information.

The active-subnetwork-oriented pathway enrichment paradigm of pathfindR can be summarized as follows: Mapping the statistical significance of each gene onto a PIN, active subnetworks, i.e., subnetworks in the PIN that contain an optimal number of significant nodes maximizing the overall significance of the subnetwork, either in direct contact or in indirect contact via an insignificant (non-input) node, are identified. Following a subnetwork filtering step, enrichment analyses are then performed on these active subnetworks. Similar to the above-mentioned PIN-aided enrichment approaches, utilization of active subnetworks allows for efficient exploitation of interaction information and enhances enrichment analysis.

For the identification of active subnetworks, various algorithms have been proposed, such as greedy algorithms (Breitling et al., 2004; Sohler et al., 2004; Chuang et al., 2007; Nacu et al., 2007; Ulitsky and Shamir, 2007; Karni et al., 2009; Ulitsky and Shamir, 2009; Fortney et al., 2010; Doungpan et al., 2016), simulated annealing (Ideker et al., 2002; Guo et al., 2007), genetic algorithms (Klammer et al., 2010; Ma et al., 2011; Wu et al., 2011; Amgalan and Lee, 2014; Ozisik et al., 2017), and mathematical programming-based methods (Dittrich et al., 2008; Zhao et al., 2008; Qiu et al., 2009; Backes et al., 2012; Beisser et al., 2012). In pathfindR, we provide implementations for a greedy algorithm, a simulated annealing algorithm, and a genetic algorithm.

In summary, pathfindR integrates information from three main resources to enhance determination of the mechanisms underlying a phenotype: (i) differential expression/methylation information obtained through omics analyses, (ii) interaction information through a PIN via active subnetwork identification, and (iii) pathway/gene set annotations from sources such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa et al., 2017), Reactome (Fabregat et al., 2018), BioCarta (Nishimura, 2001), and Gene Ontology (GO) (Ashburner et al., 2000).

The pathfindR R (https://www.R-project.org/) package was developed based on a previous approach developed by our group for genome-wide association studies (GWASes): Pathway and Network-Oriented GWAS Analysis (PANOGA) (Bakir-Gungor et al., 2014). PANOGA was successfully applied to uncover the underlying mechanisms in GWASes of various diseases, such as intracranial aneurysm (Bakir-Gungor and Sezerman, 2013), epilepsy (Bakir-Gungor et al., 2013), and Behcet's disease (Bakir-Gungor et al., 2015). With pathfindR, we aimed to extend the approach of PANOGA to omics analyses and provide novel functionality.

In this article, we present an overview of pathfindR, example applications on three gene expression data sets, and comparison of the results of pathfindR with those obtained using three tools widely used for enrichment analyses: The Database for

Annotation, Visualization and Integrated Discovery (DAVID) (Huang da et al., 2009), Signaling Pathway Impact Analysis (SPIA) (Tarca et al., 2009), and Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005).

MATERIAL AND METHODS

PINs and Gene Sets

PIN data available in pathfindR by default are KEGG, Biogrid (Stark et al., 2006; Chatr-Aryamontri et al., 2017), GeneMania (Warde-Farley et al., 2010), and IntAct (Orchard et al., 2014). The default PIN is Biogrid. Besides these four default PINs, the researcher can also use any other PIN of their choice on the condition that they provide the PIN file in simple interaction file (SIF) format.

The KEGG Homo sapiens PIN was created by an in-house script using the KEGG pathways. In KEGG, pathways are represented in XML files that contain genes and gene groups, such as protein complexes as entries and interactions as entry pairs. The KEGG pathway XML files were obtained using the official KEGG Application Programming Interface (API) which is a REST-style interface to the KEGG database resource. Using the in-house script, the XML files were parsed; the interactions were added as undirected pairs, while interaction types were disregarded. In cases of an entry in an interacting pair containing multiple genes, interactions from all of these genes to the other entry were built.

For Biogrid, Homo sapiens PIN data in tab-delimited text format from release 3.4.156 (BIOGRID-ORGANISM-Homo_sapiens-3.4.156.tab.txt) was obtained from the Biogrid Download File Repository (https://downloads.thebiogrid.org/BioGRID).

For IntAct, the PIN data in Proteomics Standards Initiative – Molecular Interactions tab-delimited (PSI-MI TAB) (MITAB) format (intact.txt) were obtained from the IntAct Molecular Interaction Database FTP site (ftp://ftp.ebi.ac.uk/pub/databases/intact/current) in January 2018.

For GeneMania, Homo sapiens PIN data in tab-delimited text format from the latest release (COMBINED.DEFAULT_NETWORKS.BP_COMBINING.txt) was obtained from the official data repository (http://genemania.org/data/current/Homo_sapiens.COMBINED/). For this PIN only, only interactions with GeneMania weights ≥0.0006 were kept, allowing only strong interactions.

No filtration for interaction types were performed for any PIN (i.e., all types of interactions were kept). The processing steps performed for all the PINs were (1.) if the HUGO Gene Nomenclature Committee (HGNC) symbols for interacting genes were not provided, conversion of provided gene identifiers to HGNC symbols using biomaRt (Durinck et al., 2009) was performed; (2.) duplicate interactions and self-interactions (if any) were removed; and (3.) all PINs were formatted as SIFs.

Gene sets available in pathfindR are KEGG, Reactome, BioCarta, GO-Biological Process (GO-BP), GO-Cellular Component (GO-CC), GO-Molecular Function (GO-MF) and GO-All (GO-BP, GO-CC, and GO-MF combined).

KEGG gene sets were obtained using the R package KEGGREST. Reactome gene sets in Gene Matrix Transposed (GMT) file format were obtained from the Reactome website (https://reactome.org/download/current/). BioCarta gene sets in GMT format were retrieved from the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) website (http://software.broadinstitute.org/gsea/msigdb). All "High-quality" GO gene sets were obtained from GO2MSIG (Powell, 2014) web interface (http://www.go2msig.org/cgi-bin/prebuilt.cgi?taxid=9606) in GMT format. All of the datasets were processed using R to obtain (1) a list containing the genes involved in each given gene set/pathway (hence, each element of the list is named by the gene set ID and is a vector of gene symbols located in the given gene set/pathway) and (2) a list containing the descriptions for each gene set/pathway (i.e., a list linking gene set IDs to description).

All of the gene sets in pathfindR are for Homo sapiens, and the default gene set is KEGG. The researcher can also use a gene set of their choice following the instructions on pathfindR wiki.

All of the default data for PINs and gene sets are planned to be updated annually.

Scoring of Subnetworks

In pathfindR, we followed the scoring scheme that was proposed by Ideker et al., 2002). The p value of each gene is converted to a z score using equation (1), and the score of a subnetwork is calculated using equation (2). In equation (1) Φ^{-1} is the inverse normal cumulative distribution function. In equation (2), A is the set of genes in the subnetwork and k is its cardinality.

$$z_i = \Phi^{-1}(1 - p_i) \tag{1}$$

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \tag{2}$$

In the same scoring scheme, a Monte Carlo approach is used for the calibration of the scores of subnetworks against a background distribution. Using randomly selected genes, 2,000 subnetworks of each possible size are constructed, and for each possible size, the mean and standard deviation of the score is calculated. These values are used to calibrate the subnetwork score using equation (3).

$$s_A = \frac{(z_A - \mu_k)}{\sigma_k} \tag{3}$$

Active Subnetwork Search Algorithms

Currently, there are three algorithms implemented in the pathfindR package for active subnetwork search, described below.

Greedy Algorithm

Greedy algorithm is the problem-solving/optimization concept that chooses locally the best option in each stage with the expectation of reaching the global optimum. In active subnetwork search, this is generally applied by starting with a significant

seed node and considering addition of a neighbor in each step to maximize the subnetwork score. In pathfindR, we used the approach described by Chuang et al. (2007). This algorithm considers addition of a node within a specified distance d to the current subnetwork. In our method, the maximum depth from the seed can also be set. With the default parameters, our greedy method considers addition of direct neighbors (d = 1) and forms a subnetwork with a maximum depth of 1 for each seed. Because the expansion process runs for each significant seed node, several overlapping subnetworks emerge. Overlapping subnetworks are handled by discarding a subnetwork that overlaps with a higher scoring subnetwork more than a given threshold, which is set to 0.5 by default.

Simulated Annealing Algorithm

Simulated annealing is an optimization algorithm inspired by annealing in metallurgy. In the annealing process, the material is heated above its recrystallization temperature and cooled slowly, allowing atoms to diffuse within the material and decrease dislocations. Analogous to this process, simulated annealing algorithm starts with a "high temperature" in which there is a high probability of accepting a solution that is worse than the current one as the solution space is explored. The acceptability of worse solutions allows a global search and escaping from local optima. The equation connecting temperature and probability of accepting a new solution is given in equation (4). In this equation, P(Acceptance) is the probability of accepting the new solution. In $score_{new}$ and $score_{current}$ are the scores of the new and the current solutions, respectively. Finally, temperature is the current temperature.

$$P\left(Acceptance\right) = \begin{cases} 1, if\ \ Score_{new} - Score_{current} > 0 \\ e^{\frac{Score_{new} - Score_{current}}{temperature}}, \quad otherwise \end{cases} \tag{4}$$

A less worse solution and higher temperature are the conditions that increase the chance of acceptation of a new solution. The probability of accepting a non-optimal action decreases in each iteration, as the temperature decreases in each step.

Simulated annealing provides improved performance over the greedy search by accepting non-optimal actions to increase exploration in the search space. In the active subnetwork search context, the search begins with a set of randomly chosen genes (the chosen genes are referred to as genes in "on" state and the not chosen genes are referred to as genes in "off" state). Connected components in this candidate solution are found, and the scores are calculated. In each iteration, the state of a random node is changed from on to off and vice versa. Connected components are found in the new solution, and their scores are calculated. If the score improves, the change is accepted. If the score decreases, the change is accepted with a probability proportional to the temperature parameter that decreases in each step.

Genetic Algorithm

Genetic algorithm is a bio-inspired algorithm that mimics evolution by implementing natural selection, chromosomal crossover, and mutation. The main phases of the genetic algorithm are "the selection phase" and "the crossover phase."

In the selection phase, parents from the existing population are selected through a fitness-based process to breed a new generation. Common selection methods are (i) roulette wheel selection in which a solution's selection probability is proportional to its fitness score, (ii) rank selection in which a solution's selection probability is proportional to its rank, thus preventing the domination of a high fitness solution to the rest, and (iii) tournament selection in which parents are selected among the members of randomly selected groups of solutions, thus giving more chance to small fitness solutions that would have little chance in other selection methods.

In the crossover phase, encoded solution parameters of the parents are exchanged analogous to chromosomal crossover. The common crossover operators are (i) single-point crossover in which the segment next to a randomly chosen point in the solution representation is substituted between parents, (ii) two-point crossover in which the segment between two randomly chosen points is substituted, and (iii) uniform crossover in which each parameter is randomly selected from either of the parents. Mutation is the process of randomly changing parameters in the offspring solutions in order to maintain genetic diversity and explore search space.

In our genetic algorithm implementation, candidate solutions represent the on/off state of each gene. In the implementation, we used rank selection and uniform crossover. In each iteration, the fittest solution of the previous population is preserved if the highest score of the current population is less than the previous population's score. In every 10 iterations, the worst scoring 10% of the population is replaced with random solutions. Because uniform cross-over and addition of random solutions make adequate contribution to the exploration of the search space, mutation is not performed under the default settings.

Selecting the Active Subnetwork Search Algorithm

The default search method in pathfindR is greedy algorithm with a search depth of 1 and maximum depth of 1. This method stands out with its simplicity and speed. This is also the "local subnetwork approach" used in the Local Enrichment Analysis (LEAN) method (Gwinner et al., 2017). As mentioned in the LEAN study, the number of subnetworks to be identified typically increases exponentially with increasing number of genes in the PIN, and the "local subnetwork approach" enables iterating over each local subnetwork and determining phenotype-related clusters. Greedy algorithm with search depth and maximum depth equal to 2 or more lets the search algorithm look further in the network for another significant gene to add to the cluster, but this may result in a slower runtime and a loss in interpretability.

Simulated annealing and genetic algorithms are heuristic methods that do not make any assumptions on the active subnetwork model. They can let insignificant genes between two clusters of significant genes to create a single connected active subnetwork. Thus, these algorithms may result in a large highest scoring active subnetwork, while the remaining subnetworks identified become small and therefore uninformative. This

tendency towards large subnetworks was attributed to a statistical bias prevalent in many tools (Nikolayeva et al., 2018).

The default active search method (greedy algorithm with a search depth of 1 and maximum depth of 1) in pathfindR was preferred because multiple active subnetworks are used for enrichment analyses. If the researcher decides to use the single highest scoring active subnetwork for the enrichment process, they are encouraged to consider greedy algorithm with greater depth, simulated annealing, or genetic algorithm.

Active-Subnetwork-Oriented Pathway Enrichment Analysis

The overview of the active-subnetwork-oriented pathway enrichment approach is presented in **Figure 1A**.

The required input is a two- or three-column table: Gene symbols, change values as log-fold change (optional) and adjusted p values associated with the differential expression/methylation data.

Initially, the input is filtered so that all p values are less than or equal to the given threshold (default is 0.05). Next, gene symbols that are not in the PIN are identified. If aliases of these gene symbols are found in the PIN, these symbols are converted to the corresponding aliases.

The processed data are then used for active subnetwork search. The identified active subnetworks are filtered via the following criteria: (i) has a score larger than the given quantile threshold (default is 0.80) and (ii) contains at least a specified number of input genes (default is 10).

For each filtered active subnetwork, using the genes contained in each of these subnetworks, separate pathway enrichment analyses are performed via one-sided hypergeometric testing. The enrichment tests use the genes in the PIN as the gene pool (i.e., background genes). Using the genes in the PIN instead of the whole genome is more appropriate and provides more statistical strength because active subnetworks are identified using only the genes in the PIN. Next, the p values obtained from the enrichment tests are adjusted (default is by Bonferroni method. However, the researcher may choose another method they prefer). Pathways with adjusted p values larger than the given threshold (default is 0.05) are discarded. These significantly enriched pathways per all filtered subnetworks are then aggregated by keeping only the lowest adjusted p value for each pathway if a pathway was found to be significantly enriched in the enrichment analysis of more than one subnetwork.

This process of active subnetwork search and enrichment analysis (active subnetwork search, filtering of subnetworks, enrichment analysis on each filtered subnetwork, and aggregation

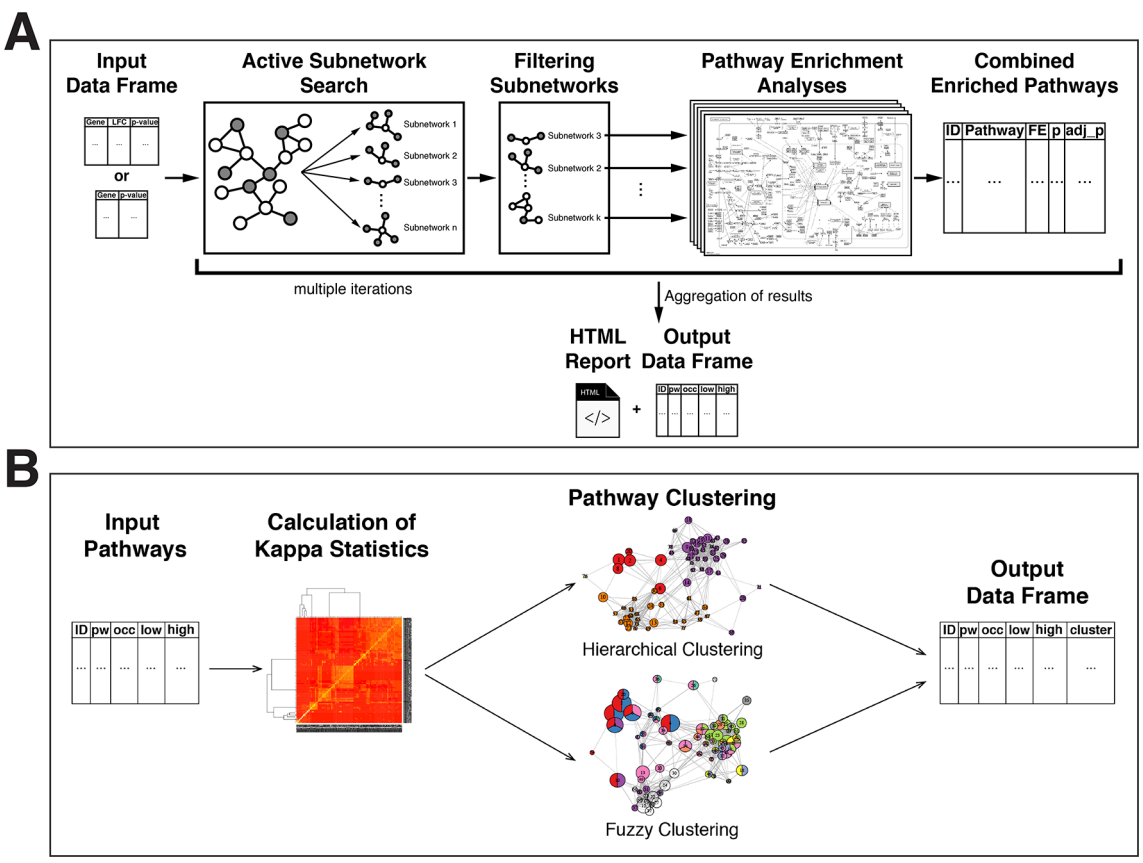


FIGURE 1 | Flow diagrams of the pathfindR methods. **(A)** Flow diagram of the pathfindR active-subnetwork-oriented pathway enrichment analysis approach. **(B)** Flow diagram of the pathfindR pathway clustering approaches.

of enrichment results over all subnetworks) is repeated for a selected number of iterations (default is 10 iterations for greedy and simulated annealing algorithms, 1 for genetic algorithm).

Finally, the lowest and the highest adjusted p values, the number of occurrences over all iterations, and up-regulated and down-regulated genes in each enriched pathway are returned as a table. Additionally, a Hypertext Markup Language (HTML) format report with the pathfindR enrichment results is created. Pathways are linked to the visualizations of the pathways if KEGG gene sets are chosen. The KEGG pathway diagrams are created using the R package pathview (Luo and Brouwer, 2013). By default, these diagrams display the involved genes colored by change values, normalized between −1 and 1, on a KEGG pathway graph. If a gene set other than KEGG is chosen and visualization is required, graphs of interactions of genes involved in the enriched pathways in the chosen PIN are visualized via the R package igraph (Csardi and Nepusz, 2006).

Pathway Clustering

Enrichment analysis usually yields a large number of related pathways. In order to establish representative pathways among similar groups of pathways, we propose that clustering can be performed either via hierarchical clustering (default) or via a fuzzy clustering method as described by Huang et al. (2007). These clustering approaches are visually outlined in **Figure 1B** and described below:

Firstly, using the input genes in each pathway, a kappa statistics matrix containing the pairwise kappa statistics, a chance-corrected measure of co-occurrence between two sets of categorized data, between the pathways is calculated (Huang et al., 2007).

By default, the wrapper function for pathway clustering, cluster_pathways, performs agglomerative hierarchical clustering (defining the distance as 1 − kappa statistic), automatically determines the optimal number of clusters by maximizing the average silhouette width, and returns a table of pathways with cluster assignments.

Alternatively, the fuzzy clustering method, previously proposed and described in detail by Huang et al. (2007), can be used to obtain fuzzy cluster assignments. Hence, this fuzzy approach allows a pathway to be a member of multiple clusters.

Finally, the representative pathway for each cluster is assigned as the pathway with the lowest adjusted p value.

Pathway Scoring Per Sample

The researcher can get an overview of the alterations of genes in a pathway via the KEGG pathway graph. To provide even more insight into the activation/repression statuses of pathways per each sample, we devised a simple scoring scheme that aggregates gene-level values to pathway scores, described below.

For an experiment values matrix (e.g., gene expression values matrix), EM, where columns indicate samples and rows indicate genes, the gene score GS of a gene g in a sample s is calculated as:

$$GS(g,s) = \frac{EM_{g,s} - \bar{X}_g}{sd_g} \quad (5)$$

Here, $\bar{X}_g$ is the mean value for gene g across all samples, and sd_g is the standard deviation for gene g across all samples.

For a set P_i, the set of k genes in pathway i, and a sample j, the i^{th} row and j^{th} column of the pathway score matrix PS is calculated as follows:

$$PS_{i,j} = \frac{1}{k}\sum_{g \in P_i} GS(g, j) \quad (6)$$

The pathway score of a sample for a given pathway is therefore the average value of the scores of the genes in the pathway for the given sample.

After calculation of the pathway score matrix, a heat map of these scores is plotted. Via this heat map, the researcher can examine the activity of a pathway in individual samples as well as compare the overall activity of the pathway between cases and controls.

Application on Gene Expression Datasets

To analyze the performance of pathfindR, we used three gene expression datasets. All datasets were obtained via the Gene Expression Omnibus (GEO) (Edgar et al., 2002). The first dataset (GSE15573) aimed to characterize and compare gene expression profiles in the peripheral blood mononuclear cells of 18 rheumatoid arthritis (RA) patients versus 15 healthy subjects using the Illumina human-6 v2.0 expression bead chip platform. This dataset will be referred to as RA. The second dataset (GSE4107) compared the gene expression profiles of the colonic mucosa of 12 early onset colorectal cancer patients and 10 healthy controls using the Affymetrix Human Genome U133 Plus 2.0 Array platform. The second dataset will be referred to as CRC. The third dataset (GSE55945) compared the expression profiles of prostate tissue from 13 prostate cancer patients versus 8 controls using the Affymetrix Human Genome U133 Plus 2.0 Array platform. This dataset will be referred to as PCa.

After preprocessing, which included log2 transformation and quantile normalization, differential expression testing via a moderated t test using limma (Ritchie et al., 2015) was performed. Next, the resulting p values were corrected using false discovery rate (FDR) adjustment. The differentially expressed genes (DEGs) were defined as those with FDR < 0.05. Probes mapping to multiple genes and probes that do not map to any gene were excluded. If a gene was targeted by multiple probes, the lowest p value was kept. The results of differential expression analyses for RA, CRC, and PCa, prior to filtering (differential expression statistics for all probes) and after filtering (lists of DEGs), are provided in **Supplementary Data Sheet 1**.

We chose to use these three datasets because these are well-studied diseases and the involved mechanisms are considerably well characterized. These different datasets also allowed us to test the capabilities of pathfindR on DEGs obtained from different platforms.

We performed enrichment analysis with pathfindR, using the default settings. Greedy algorithm for active subnetwork search was used, and the analysis was carried out over 10 iterations. The enrichment significance cutoff value was set to 0.25 for each

analysis (changing the argument enrichment_threshold of run_pathfindR function) as we later performed validation of the results using the three significance cutoff values of 0.05, 0.1, and 0.25

To better evaluate the performance of pathfindR, we compared results on the three gene expression datasets by three widely used pathway analysis tools, namely, DAVID (Huang da et al., 2009), SPIA (Tarca et al., 2009), and GSEA (Subramanian et al., 2005). DAVID 6.8 was used for the analyses. SPIA was performed using the default settings. GSEA was also performed using the default settings (using phenotype permutations). Additionally, pre-ranked GSEA was performed (GSEAPreranked) using the default settings. The rank of the i^{th} gene $rank_i$ was calculated as follows:

$$rank_i = \begin{cases} -p_i^{-1}, if\ \ logFC_i < 0 \\ p_i^{-1}, otherwise \end{cases} \quad (7)$$

The unfiltered results of enrichment analyses using the different methods on the three datasets are presented in **Supplementary Data Sheet 2**.

For each analysis, the Bonferroni-corrected p values for pathfindR were used to filter the results. For all the other tools, as the Bonferroni method would be too strict and result in too few or no significant pathways, the FDR-corrected p values were used.

Because there is no definitive answer to which pathways are involved in the pathogenesis of the conditions under study, we analyzed the results in light of the existing biological knowledge on the conditions and compared our results with other tools in this context. The significant pathways were assessed on the basis of how well they fitted with the existing knowledge. For this, two separate approaches were taken: (i) assessment of literature support for the significantly enriched pathways (using a significance threshold of 0.05), and (ii) assessment of the percentages of pathway genes that are also known disease genes (using the three significance thresholds of 0.05, 0.10, and 0.25). While both assessments could be separately used to determine the "disease-relatedness" of a pathway, we chose to use them both as these are complementary measures: the former is a more subjective but a comprehensive measure of association, and the latter is a limited but a more objective measure of association. For determining the percentages of known disease genes in each significantly enriched pathway, two curated lists were used. For the RA dataset, mapped genes in the curated list of SNPs associated with RA was obtained from the NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog, retrieved on 19.12.2018) (MacArthur et al., 2017). These genes will be referred to as "RA Genes." For the CRC and PCa datasets, the "Cancer Gene Census" (CGC) genes from the Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk, retrieved on 19.12.2018) were used. These genes will be referred to as "CGC Genes."

Assessment Using Permuted Inputs

We performed pathfindR analyses using real and permuted data with different sizes to assess the number of enriched pathways identified in the permuted data against the actual data. For this assessment, the RA data was used. The analyses were performed on data subsets taken as the top 200, 300, 400, and 500 most significant DEGs as well as the complete list of 572 DEGs. For each input size, 100 separate pathfindR analyses were performed on both the actual input data and permuted data. While the real input data were kept unchanged, for the permuted data, a random permutation of genes (using the set of all genes available on the microarray platform) was carried out at each iteration over 100 analyses. Analyses with pathfindR were performed using the default settings described above.

The distributions of the number of enriched pathways for actual vs. permuted data were compared using Wilcoxon rank sum test.

ORA Assessment of the Effect of DEGs Without Any Interactions

We performed ORA as implemented in pathfindR (as the "enrichment" function) to assess any effect of removing DEGs without any interactions on enrichment results. For this purpose, ORA were performed for (i) the full lists of DEGs for all datasets and (ii) the lists of DEGs that are found in the Biogrid PIN. As gene sets, KEGG pathways were used. As background genes, all of the genes in the Biogrid PIN were used for both analyses so that the results could be comparable. The enrichment p values were adjusted using the FDR method. Pathway enrichment was considered significant if FDR was <0.05.

Assessment of the Effect of PINs on Enrichment Results

To analyze the effect of the chosen PIN on the enrichment results, we performed pathfindR analyses using the four PINs provided by default: the Biogrid, GeneMania, IntAct, and KEGGPINs. For these analyses, the default settings were used with the default active search algorithm (greedy) and the default gene sets (KEGG).

Software Availability

The pathfindR package is freely available for use under MIT license: https://cran.r-project.org/package=pathfindR. The code of the pathfindR package is deposited in a GitHub repository (https://github.com/egeulgen/pathfindR) along with a detailed wiki, documenting the features of pathfindR in detail. Docker images for the latest stable version and the development version of pathfindR are deposited on Docker Hub (https://hub.docker.com/r/egeulgen/pathfindr)

RESULTS

The RA Dataset

A total of 572 DEGs were identified for the RA dataset (**Supplementary Data Sheet 1**). Filtered by adjusted p values (adjusted-p ≤ 0.05), pathfindR identified 78 significantly enriched KEGG pathways which were partitioned into 10 clusters

(**Figures 2A**, **B**). The relevancy of 31 out of 78 (39.74%) pathways was supported by literature, briefly stated in **Table 1**.

The summary of results obtained using the different tools and literature support for the identified pathways (where applicable) are presented in **Table 1**. For this dataset, SPIA identified two significant pathways, which were both also identified by pathfindR. No significant pathway was identified by the other tools.

Clustering allowed us to obtain coherent groups of pathways and identify mechanisms relevant to RA, including autoimmune response to the spliceosome (Hassfeld et al., 1995), mechanisms related with response to microbial infection, such as generation of neo-autoantigens and molecular mimicry (Li et al., 2013), dysregulation of various signaling pathways (Remans et al., 2002; Rihl et al., 2005; Barthel et al., 2009; Malemud, 2015), DNA damage repair (Lee et al., 2003), dysregulation of energy metabolism (Yang et al., 2015), and modulation of immune response and inflammation by the proteasome (Wang and Maldonado, 2006).

The activity scores of the representative pathways for each subject indicated that most representative pathways were down-regulated in the majority of subjects (**Figure 2C**).

The CRC Dataset

For the CRC dataset, 1,356 DEGs were identified (**Supplementary Data Sheet 1**). pathfindR identified 100 significantly enriched pathways (adjusted-p ≤ 0.05) which were partitioned into 14 coherent clusters (**Figures 3A**, **B**). Forty-eight (48%) of these enriched pathways were relevant to CRC biology, as supported by literature. Brief descriptions of how these are relevant are provided in **Table 2**.

The results obtained using the different tools and literature support for the identified pathways (where applicable) are presented in **Table 2**. For this dataset, DAVID identified 20 significant pathways, 15 of which were also found by pathfindR (4 out of the remaining 5 were not supported by literature to be relevant to CRC). SPIA identified 13 significantly enriched pathways, 11 of which were also identified by pathfindR. Out of the remaining two enriched pathways, only "PPAR signaling pathway" was related to CRC biology (You et al., 2015). Neither GSEA nor GSEAPreranked yielded any significant pathways for the CRC dataset. The Colorectal cancer pathway was identified to be significantly enriched only by pathfindR.

Upon clustering, 14 clusters were identified (**Figures 3A**, **B**). These clusters implied processes previously indicated

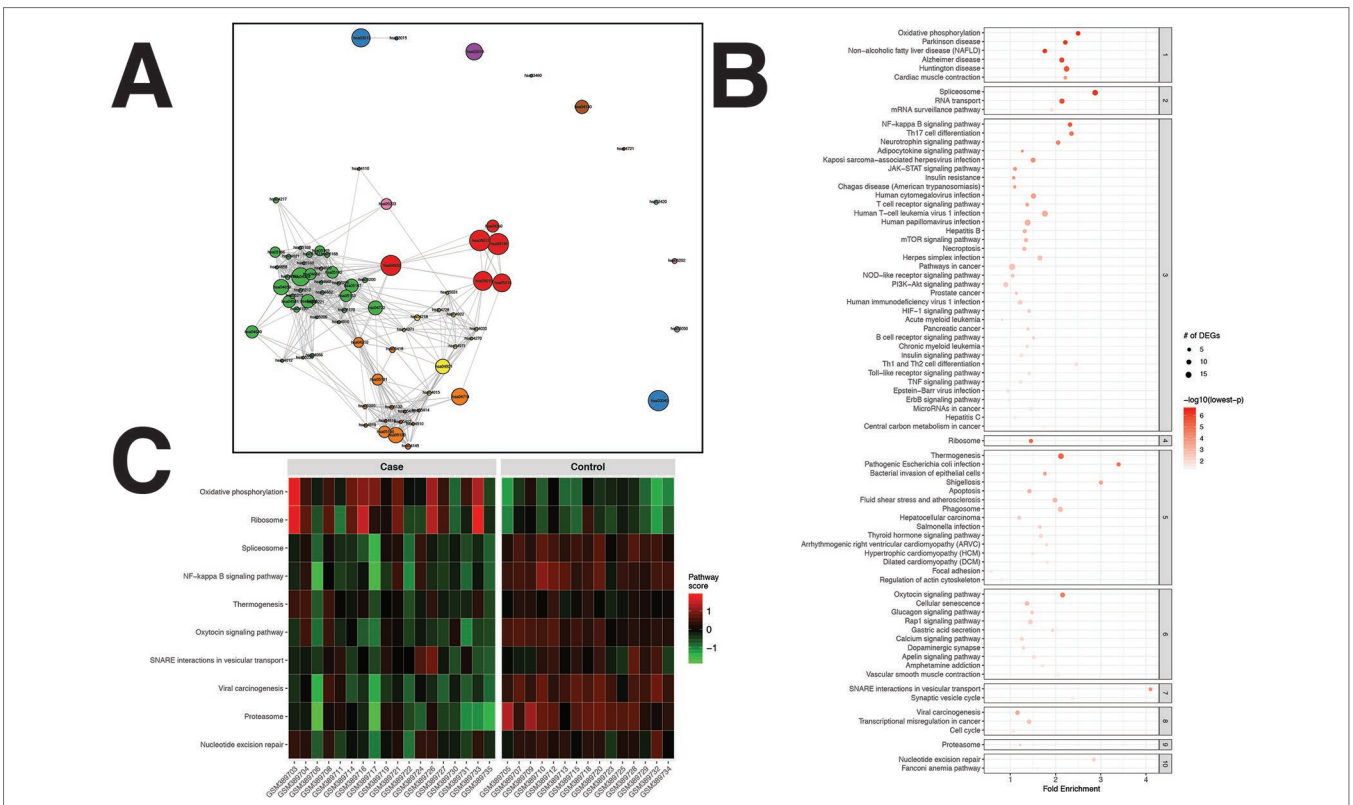


FIGURE 2 | pathfindR enrichment and clustering results on the rheumatoid arthritis (RA) dataset (lowest p ≤ 0.05). **(A)** Clustering graph, each color displaying the clusters obtained for RA. Each node is an enriched pathway. Size of a node corresponds to its –log(lowest_p). The thickness of the edges between nodes corresponds to the kappa statistic between the two terms. **(B)** Bubble chart of enrichment results grouped by clusters (labeled on the right-hand side of each panel). The x axis corresponds to fold enrichment values, while the y axis indicates the enriched pathways. The size of the bubble indicates the number of differentially expressed genes (DEGs) in the given pathway. Color indicates the –log10(lowest-p) value; the more it shifts to red, the more significantly the pathway is enriched. **(C)** Heat map of pathway scores per subject. The x axis indicates subjects, whereas the y axis indicates representative pathways. Color scale for the pathway score is provided in the right-hand legend.

TABLE 1 | Pathway analysis results for the rheumatoid arthritis (RA) dataset (adjusted $p < 0.05$).

ID	Pathway	% RA genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa00190	Oxidative phosphorylation	0	**<0.001**	0.28157363	–	0.50656915	1	Oxygen metabolism has an important role in the pathogenesis of RA (Hitchon and El-Gabalawy, 2004; Yang et al., 2015).
hsa05012	Parkinson disease	1.41	**<0.001**	0.35202042	**0.03287527**	0.5198511	1	
hsa03040	Spliceosome	0	**<0.001**	0.19110635	–	–	–	Autoimmune response to the spliceosome was previously reported in numerous autoimmune diseases (Hassfeld et al., 1995).
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	2.01	**<0.001**	–	–	–	–	
hsa05010	Alzheimer disease	0.58	**<0.001**	0.40188326	0.070524222	0.49685246	0.99091035	
hsa03013	RNA transport	0.61	**<0.001**	0.49158247	0.080862112	–	–	
hsa05016	Huntington disease	0.52	**<0.001**	0.24543866	0.03287527	0.5436461	1	
hsa04064	NF-kappa B signaling pathway	8.42	**<0.001**	0.634065	0.206122248	–	–	NF-kB is a pivotal mediator of inflammatory responses (Liu et al., 2017b) and an important player in RA pathogenesis (Makarov, 2001).
hsa03010	Ribosome	0	**<0.001**	–	–	0.6974111	–	
hsa04714	Thermogenesis	0.43	**<0.001**	–	–	–	–	
hsa05130	Pathogenic Escherichia coli infection	0	**<0.001**	0.42959791	0.103834432	0.740603	0.96458197	Possibly related to generation of neo-autoantigens, molecular mimicry, and bystander activation of the immune system (Li et al., 2013)
hsa04659	Th17 cell differentiation	19.63	**<0.001**	–	–	–	–	Th17 cells play an important role in inflammation in human autoimmune arthritides, including RA (Pernis, 2009; Leipe et al., 2010).
hsa04921	Oxytocin signaling pathway	1.97	**<0.001**	–	–	–	–	
hsa04722	Neurotrophin signaling pathway	2.52	**<0.001**	0.55824289	0.331277414	–	–	Neurotrophin signaling is altered in RA (Rihl et al., 2005; Barthel et al., 2009).
hsa04130	SNARE interactions in vesicular transport	0	**<0.001**	0.51353532	0.205302976	0.69465846	0.9727782	
hsa04920	Adipocytokine signaling pathway	2.9	**<0.001**	–	0.999995202	–	–	The adipocytokines and the adipokine network have extensive roles in the pathogenesis of RA (Frommer et al., 2011; Del Prete et al., 2014).
hsa05167	Kaposi sarcoma-associated herpes virus infection	5.91	**<0.001**	–	–	–	–	
hsa04630	JAK-STAT signaling pathway	9.26	**<0.001**	–	0.980050749	–	–	Disruption of the JAK-STAT pathway is a critical event in the pathogenesis and progression of rheumatoid arthritis (Malemud, 2018).

(*Continued*)

TABLE 1 | Continued

ID	Pathway	% RA genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04931	Insulin resistance	1.85	**<0.001**	–	–	–	–	
hsa04260	Cardiac muscle contraction	0	**<0.001**	–	–	0.6976311	1	
hsa05142	Chagas disease (American trypanosomiasis)	7.77	**<0.001**	–	0.999995202	–	–	
hsa05100	Bacterial invasion of epithelial cells	1.35	**<0.001**	–	0.743380146	–	–	Possibly related to generation of neo-autoantigens, molecular mimicry, and bystander activation of the immune system (Li et al., 2013).
hsa05163	Human cytomegalovirus infection	4	**<0.001**	–	–	–	–	
hsa04660	T cell receptor signaling pathway	10.89	**<0.001**	–	0.743380146	–	–	Dysregulation of the TCR signaling pathway was previously implicated in RA biology (Sumitomo et al., 2018).
hsa05131	Shigellosis	3.08	**<0.001**	0.51130292	0.137642182	–	–	Possibly related to generation of neo-autoantigens, molecular mimicry, and bystander activation of the immune system (Li et al., 2013).
hsa05203	Viral carcinogenesis	2.99	**<0.001**	–	0.999995202	–	–	
hsa05166	Human T-cell leukemia virus 1 infection	7.31	**<0.001**	0.48795724	0.137642182	–	–	
hsa04210	Apoptosis	1.47	**<0.001**	–	0.827952041	–	–	Apoptosis may play divergent roles in RA biology (Liu and Pope, 2003).
hsa05165	Human papillomavirus infection	1.82	**<0.001**	–	–	–	–	
hsa05161	Hepatitis B	6.13	**<0.001**	–	–	–	–	
hsa04150	mTOR signaling pathway	0.66	**0.001061744**	–	0.743380146	–	–	Intracellular signaling pathway (including mTOR signaling) play a critical role in rheumatoid arthritis (Malemud, 2013; Malemud, 2015).
hsa05418	Fluid shear stress and atherosclerosis	1.44	**0.001166905**	–	–	–	–	
hsa04218	Cellular senescence	2.5	**0.001351009**	–	–	–	–	
hsa04217	Necroptosis	4.32	**0.001442161**	–	–	–	–	Necroptosis suppresses inflammation via termination of TNF- or LPS-induced cytokine and chemokine production (Kearney et al., 2015).
hsa04145	Phagosome	5.26	**0.001665316**	0.49641734	–	–	–	
hsa03050	Proteasome	2.22	**0.001881322**	–	–	0.7889826	–	Proteasome modulates immune and inflammatory responses in autoimmune diseases (Wang and Maldonado, 2006).
hsa05168	Herpes simplex infection	8.65	**0.002442405**	–	0.53977679	–	–	

(*Continued*)

TABLE 1 | Continued

ID	Pathway	% RA genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa05200	Pathways in cancer	4.56	**0.002463658**	–	0.743380146	–	–	
hsa04621	NOD-like receptor signaling pathway	5.06	**0.002477183**	–	0.909381246	–	–	NOD-like receptors are being implicated in the pathology of RA and other rheumatic diseases (McCormack et al., 2009).
hsa05202	Transcriptional misregulation in cancer	4.84	**0.002495122**	–	0.743380146	–	–	
hsa04151	PI3K-Akt signaling pathway	2.82	**0.00331152**	–	–	–	–	PI3K-Akt signaling regulates diverse cellular processes and was proposed as a target for inducing cell death in RA (Malemud, 2015).
hsa05215	Prostate cancer	1.03	**0.003884234**	–	0.999995202	–	–	
hsa05170	Human immunodeficiency virus 1 infection	3.3	**0.004185672**	–	–	–	–	
hsa04066	HIF-1 signaling pathway	5	**0.004382877**	–	–	–	–	Alterations in hypoxia-related signaling pathways are considered potential mechanisms of RA pathogenesis (Quiñonez-Flores et al., 2016).
hsa05225	Hepatocellular carcinoma	3.57	**0.004782642**	–	–	–	–	
hsa04922	Glucagon signaling pathway	0	**0.004927201**	–	–	–	–	
hsa03420	Nucleotide excision repair	0	**0.005418059**	0.63260927	–	–	–	DNA damage load is higher in RA patients, thus activating repair pathways (Lee et al., 2003).
hsa04015	Rap1 signaling pathway	0.97	**0.005543915**	–	–	–	–	Deregulation of Rap1 signaling pathway was shown to be a critical event altering the response of synovial T cells in RA (Remans et al., 2002).
hsa05221	Acute myeloid leukemia	3.03	**0.006008327**	–	0.999995202	–	–	
hsa05132	Salmonella infection	3.49	**0.006557353**	–	0.721697645	–	–	Possibly related to generation of neo-autoantigens, molecular mimicry, and bystander activation of the immune system (Li et al., 2013).
hsa05212	Pancreatic cancer	4	**0.006646458**	–	0.743380146	–	–	
hsa04662	B cell receptor signaling pathway	2.82	**0.00748718**	–	0.851804025	–	–	
hsa04971	Gastric acid secretion	4	**0.008829291**	–	0.743380146	–	–	
hsa04020	Calcium signaling pathway	3.19	**0.009304653**	–	0.999995202	–	–	Dysregulation of the calcium signaling pathway was implicated in RA pathogenesis (Berridge, 2016).
hsa04919	Thyroid hormone signaling pathway	3.45	**0.009478272**	–	–	–	–	
hsa05220	Chronic myeloid leukemia	3.95	**0.011899647**	–	0.743380146	–	–	
hsa04728	Dopaminergic synapse	1.53	**0.012701709**	–	0.743380146	–	–	

(*Continued*)

TABLE 1 | Continued

ID	Pathway	% RA genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1.39	**0.012829444**	–	0.96639695	–	–	
hsa04371	Apelin signaling pathway	1.46	**0.015134748**	–	–	–	–	
hsa04910	Insulin signaling pathway	0	**0.015134748**	–	0.999995202	–	0.95215786	
hsa03015	mRNA surveillance pathway	0	**0.015767824**	–	–	–	–	
hsa04658	Th1 and Th2 cell differentiation	17.39	**0.016291789**	–	–	–	–	RA patients were characterized by a disruption of Th1/Th2 balance towards Th1(He et al., 2017).
hsa04620	Toll-like receptor signaling pathway	5.77	**0.017712009**	–	0.743380146	0.72964895	1	Toll-like receptors are being implicated in the pathology of RA and other rheumatic diseases (McCormack et al., 2009).
hsa05410	Hypertrophic cardiomyopathy (HCM)	2.41	**0.019642576**	–	–	–	–	
hsa04668	TNF signaling pathway	5.45	**0.0209396**	–	–	–	–	Intracellular signaling pathway (including TNF signaling) play a critical role in rheumatoid arthritis (Malemud, 2013).
hsa05169	Epstein-Barr virus infection	8.96	**0.022925676**	–	0.743380146	–	–	
hsa05031	Amphetamine addiction	2.94	**0.023901842**	–	0.743380146	–	–	
hsa05414	Dilated cardiomyopathy (DCM)	2.22	**0.025016113**	–	0.851804025	–	–	
hsa04012	ErbB signaling pathway	2.35	**0.026253837**	–	0.999995202	–	–	Intracellular signaling pathway play a critical role in rheumatoid arthritis (Malemud, 2013).
hsa04510	Focal adhesion	0.5	**0.02805129**	–	0.999995202	0.77597433	–	Adhesion molecules have an important role in RA (Pitzalis et al., 1994).
hsa04110	Cell cycle	4.03	**0.029916503**	–	0.743380146	–	–	Cell cycle stalling was recently linked to arthritis (Matsuda et al., 2017).
hsa05206	MicroRNAs in cancer	1.34	**0.03026234**	–	–	–	–	
hsa03460	Fanconi anemia pathway	0	**0.033094195**	–	0.743380146	–	–	DNA damage load is higher in RA patients, thus activating repair pathways (Lee et al., 2003).
hsa05160	Hepatitis C	3.23	**0.035219047**	–	0.743380146	–	–	
hsa04721	Synaptic vesicle cycle	1.28	**0.035442941**	–	–	–	–	
hsa04810	Regulation of actin cytoskeleton	0.47	**0.036496481**	–	0.96639695	0.80830806	–	Actin cytoskeleton dynamics is linked to synovial fibroblast activation (Vasilopoulos et al., 2007). Autoimmune response to cytoskeletal proteins (including actin) was reported in RA (Shrivastav et al., 2002).
hsa04270	Vascular smooth muscle contraction	2.48	**0.036862558**	–	0.070524222	–	1	

(*Continued*)

TABLE 1 | Continued

ID	Pathway	% RA genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa05230	Central carbon metabolism in cancer	1.54	**0.038909519**	–	–	–	–	Dysregulation of energy metabolism is indicated in RA (Yang et al., 2015).

"ID" indicates the Kyoto Encyclopedia of Genes and Genomes (KEGG) ID for the enriched pathway, whereas "Pathway" indicates the KEGG pathway name. "% RA genes" indicates the percentage of RA genes in the pathway. The lowest Bonferroni-adjusted p value for pathfindR analysis is provided in "pathfindR," the false discovery rate (FDR)-adjusted p value for Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis is provided in "DAVID," the FDR-adjusted p value for Signaling Pathway Impact Analysis (SPIA) is presented in "SPIA," and the FDR-adjusted p values for Gene Set Enrichment Analysis (GSEA) and GSEAPreranked are presented in "GSEA" and "GSEAPreranked," respectively. Significant p values (i.e., adjusted p value <0.05) are given in bold font. "-" indicates the pathway was not found to be enriched by the given tool. If a pathway is relevant to RA, a brief description of its relevance is provided in "Brief Description."

in colorectal cancer, including but not limited to colorectal cancer and related signaling pathways (Fang and Richardson, 2005; Zenonos and Kyprianou, 2013; Francipane and Lagasse, 2014), apoptosis (Watson, 2004), p53 signaling (Slattery et al., 2018), dysregulation of metabolic functions, including glucose metabolism (Fang and Fang, 2016), fatty acid metabolism (Wen et al., 2017), and amino acid metabolism (Santhanam et al., 2016; Antanaviciute et al., 2017), and cell cycle (Hartwell and Kastan, 1994; Collins et al., 1997; Jarry et al., 2004). Brief descriptions of all pathways relevant to CRC are provided in **Table 2**.

Representative pathways that were upregulated in the majority of subjects included important pathways related to cancer in general and colorectal cancer, such as the proteoglycans in cancer, adherens junction, gap junction, and Hippo signaling pathway. Representative pathways that were downregulated in the majority of subjects included other important pathways related to colorectal cancer, such as valine, leucine, and isoleucine degradation, mTOR signaling pathway, and cell cycle (**Figure 3C**).

The PCa Dataset

For the PCa dataset, 1,240 DEGs were identified (**Supplementary Data Sheet 1**). pathfindR identified 92 significantly enriched pathways (adjusted-p ≤ 0.05) which were clustered into 14 coherent clusters (**Figures 4A**, **B**). Forty-six (50%) of these enriched pathways were relevant to PCa biology, as supported by literature. Brief descriptions of the relevancies are provided in **Table 3**.

The results obtained using the different tools and literature support for the identified pathways (where applicable) are presented in **Table 3**. DAVID identified eight significant pathways, which were all also identified by pathfindR and only half of which were relevant to PCa. SPIA identified five significantly enriched pathways, all of which were also identified by pathfindR. GSEA identified no significant pathways, whereas GSEAPreranked identified one significant pathway, for which no association with PCa was provided by the literature. The prostate cancer pathway was identified to be significantly enriched only by pathfindR.

The clusters identified by pathfindR pointed to several mechanisms previously shown to be important for prostate cancer. These mechanisms included but were not limited to the prostate cancer pathway and related signaling pathways (El Sheikh et al., 2003; Shukla et al., 2007; Rodríguez-Berriguete et al., 2012), cancer immunity (Knutson and Disis, 2005; Zhao et al., 2014), Hippo signaling (Zhang et al., 2015), cell cycle (Balk and Knudsen, 2008), autophagy (Farrow et al., 2014), and insulin signaling (Cox et al., 2009; Bertuzzi et al., 2016).

The majority of representative pathways relevant to PCa were down-regulated (**Figure 4C**).

Common Pathways Between the CRC and PCa Datasets

Because the CRC and PCa datasets were both cancers, they were expected to have common pathways identified by pathfindR. Indeed, 47 common significant pathways (adjusted-p ≤ 0.05) were identified (**Supplementary Table 1**). These common pathways included general cancer-related pathways, such as pathways in cancer, proteoglycans in cancer, MAPK signaling pathway, Ras signaling pathway, Hippo signaling pathway, mTOR signaling pathway, Toll-like receptor signaling pathway, Wnt signaling pathway, and adherens junction.

Disease-Related Genes in the Significantly Enriched Pathways

The percentages of disease-related genes for each pathway found to be enriched by any tool (adjusted-p ≤ 0.05) are presented in the corresponding columns of **Tables 1**, **2**, and **3** ("% RA Genes" for the RA dataset and "% CGC Genes" for the CRC and PCa datasets). These percentages show great variability but support the literature search results in assessing the disease-relatedness of the enriched pathways.

The distributions of disease-related gene percentages in pathways identified by each tool in the three different datasets, filtered by the adjusted-p value thresholds of 0.05, 0.1, and 0.25, are presented in **Figure 5**. As stated before, for the RA dataset, only pathfindR and SPIA identified significant pathways. The median percentages of RA-associated genes of the enriched pathways of pathfindR was higher than the median percentages of SPIA (2.43% vs. 0.96% for the 0.05 cutoff, 2.5% vs. 0.61% for the 0.1 cutoff, and 2.27% vs. 0.67% for the 0.25 cutoff). For CRC, pathfindR displayed the highest median percentage of CGC genes for all the cutoff values (17.84%, 17.72%, and 16.7% for 0.05, 0.1, and 0.25, respectively). For the PCa dataset, the median

TABLE 2 | Pathway analysis results for the colorectal cancer (CRC) dataset (adjusted $p < 0.05$).

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04974	Protein digestion and absorption	5.56	**<0.001**	**0.01573699**	–	–	–	
hsa04512	ECM-receptor interaction	6.1	**<0.001**	**0.00010652**	<0.001	0.3232827	0.92760116	The extracellular matrix modulates the hallmarks of cancer (Pickup et al., 2014).
hsa04380	Osteoclast differentiation	21.26	**<0.001**	–	0.418726575	–	–	
hsa05205	Proteoglycans in cancer	27.86	**<0.001**	**0.02168567**	–	–	–	Proteoglycans play roles in modulating cancer progression, invasion and metastasis (Iozzo and Sanderson, 2011).
hsa05130	Pathogenic Escherichia coli infection	10.91	**<0.001**	0.25769925	**0.015730997**	0.23110063	1	Pathogenic E. coli is claimed to be a cofactor in pathogenesis of colorectal cancer (Bonnet et al., 2014).
hsa00280	Valine, leucine and isoleucine degradation	2.08	**<0.001**	**<0.001**	–	–	–	Degradation of branched chain amino acids could play an important role in the energy supply of cancer cells (Antanaviciute et al., 2017).
hsa04010	MAPK signaling pathway	17.97	**<0.001**	0.08238577	**0.004151739**	0.28760567	1	MAPK signaling plays an important part in progression of colorectal cancer (Fang and Richardson, 2005).
hsa04520	Adherens junction	31.94	**<0.001**	0.0852993	–	0.39334586	0.98078984	Dysregulation of the adherens junction system has particular implications in transformation and tumor invasion (Knights et al., 2012).
hsa04810	Regulation of actin cytoskeleton	15.02	**<0.001**	**0.01469723**	**0.004105905**	0.31124064	1	Regulation of actin cytoskeleton is dysregulated in cancer cell migration and invasion (Yamaguchi and Condeelis, 2007).
hsa05166	Human T-cell leukemia virus 1 infection	27.4	**<0.001**	–	0.858709076	–	–	
hsa04510	Focal adhesion	19.1	**<0.001**	**<0.001**	**<0.001**	0.2348305	0.95611423	Cancer cells exhibit highly altered focal adhesion dynamics (Maziveyi and Alahari, 2017).
hsa04540	Gap junction	19.32	**<0.001**	**0.03853227**	**0.009701705**	0.24327032	0.9830453	Deficiencies in cell-to-cell communication, particularly gap junctional intercellular communication are observed in CRC (Bigelow and Nguyen, 2014).
hsa05012	Parkinson disease	6.34	**<0.001**	**0.28728621**	**0.025978875**	–	0.91026866	
hsa04662	B cell receptor signaling pathway	42.25	**<0.001**	–	0.500093708	0.27041057	1	
hsa00071	Fatty acid degradation	6.82	**<0.001**	**<0.001**	–	–	–	Adipocytes activate mitochondrial fatty acid oxidation and autophagy to promote tumor growth in colon cancer (Wen et al., 2017).

(*Continued*)

TABLE 2 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04145	Phagosome	5.26	**<0.001**	–	–	–	–	
hsa04658	Th1 and Th2 cell differentiation	19.57	**<0.001**	–	–	–	–	T helper cells are important in cancer immunity (Knutson and Disis, 2005).
hsa05165	Human papillomavirus infection	19.09	**<0.001**	–	–	–	–	
hsa05161	Hepatitis B	31.29	**<0.001**	–	–	–	–	
hsa00640	Propanoate metabolism	0	**<0.001**	<0.001	–	–	–	
hsa04151	PI3K-Akt signaling pathway	21.47	**<0.001**	0.07244833	–	–	–	PI3K-Akt signaling is deregulated in CRC (Danielsen et al., 2015; Zhang et al., 2017a).
hsa04660	T cell receptor signaling pathway	31.68	**<0.001**	–	0.698350894	0.44643503	0.965013	T-cell receptor signaling modulates control of anti-cancer immunity (Cronin and Penninger, 2007).
hsa04659	Th17 cell differentiation	26.17	**<0.001**	–	–	–	–	A unique change of Th17 cells was observed in the progression of CRC (Wang et al., 2012).
hsa04933	AGE-RAGE signaling pathway in diabetic complications	31	**<0.001**	–	–	–	–	
hsa04657	IL-17 signaling pathway	9.68	**<0.001**	–	–	–	–	IL-17 is considered as a promoter factor in CRC progression (Wu et al., 2013).
hsa04625	C-type lectin receptor signaling pathway	27.88	**<0.001**	–	–	–	–	C-Type lectin receptors may be targeted for cancer immunity (Yan et al., 2015).
hsa05167	Kaposi sarcoma-associated herpesvirus infection	24.73	**<0.001**	–	–	–	–	
hsa05170	Human immunodeficiency virus 1 infection	16.98	**<0.001**	–	–	–	–	
hsa04921	Oxytocin signaling pathway	13.16	**<0.001**	0.1513106	–	–	–	
hsa05168	Herpes simplex infection	10.81	**<0.001**	–	0.840617856	–	–	
hsa04668	TNF signaling pathway	18.18	**<0.001**	–	–	–	–	TNF-α was shown to promote colon cancer cell migration and invasion (Zhao and Zhang, 2018).
hsa04022	cGMP-PKG signaling pathway	11.04	**<0.001**	0.00352246	–	–	–	cGMP-PKG signaling inhibits cell proliferation and induces apoptosis (Fajardo et al., 2014).
hsa00650	Butanoate metabolism	0	**<0.001**	<0.001	–	–	–	Butanoate has the ability to inhibit carcinogenesis (Goncalves and Martel, 2013).
hsa05132	Salmonella infection	8.14	**<0.001**	–	0.524757851	–	–	
hsa05014	Amyotrophic lateral sclerosis (ALS)	15.69	**<0.001**	0.36800171	0.200174194	0.27973756	1	

(*Continued*)

TABLE 2 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04530	Tight junction	11.18	**<0.001**	0.0915822	0.02704172	0.27459267	0.98746127	Dysregulation of tight junctions promote tumorigenesis as well as tumor progression in colorectal cancer (Hollande and Papin, 2013).
hsa04150	mTOR signaling pathway	16.45	**<0.001**	–	0.999999998	0.31433496	1	mTOR signaling is accepted as one of the primary mechanisms for sustaining tumor outgrowth and metastasis and is dysregulated in many cancers, including colorectal cancer (Francipane and Lagasse, 2014).
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	14.71	**<0.001**	–	0.552502996	0.327181	1	
hsa05418	Fluid shear stress and atherosclerosis	18.71	**<0.001**	–	–	–	–	
hsa04015	Rap1 signaling pathway	18.93	**<0.001**	–	–	–	–	Rap1 signaling has roles in tumor cell migration and invasion (Zhang et al., 2017b).
hsa05164	Influenza A	15.79	**<0.001**	–	0.999999998	–	–	
hsa05100	Bacterial invasion of epithelial cells	22.97	**<0.001**	–	0.167771421	–	–	
hsa05146	Amebiasis	12.5	**<0.001**	–	0.418726575	–	–	
hsa00380	Tryptophan metabolism	2.5	**<0.001**	0.0036283	–	–	–	Tryptophan metabolism is a promising target for immunotherapy in CRC (Santhanam et al., 2016).
hsa04072	Phospholipase D signaling pathway	20.55	**0.001079935**	–	–	–	–	Phospholipase D signaling has roles in cell migration, invasion and metastasis (Gomez-Cambronero, 2014).
hsa04014	Ras signaling pathway	18.1	**0.001165472**	–	–	–	–	Ras signaling has roles in colorectal cancer progression, treatment response, prognosis (Zenonos and Kyprianou, 2013).
hsa05210	Colorectal cancer	51.16	**0.00129243**	–	0.177026287	0.5272962	1	The pathway of the disease.
hsa05200	Pathways in cancer	26.81	**0.001394025**	0.01610123	0.004421859	0.23207118	0.99618906	"Meta"-pathway of cancer pathways.
hsa05169	Epstein-Barr virus infection	21.39	**0.001483431**	–	0.999999998	–	–	
hsa04934	Cushing syndrome	22.73	**0.002134647**	–	–	–	–	
hsa00190	Oxidative phosphorylation	3.76	**0.002190056**	–	–	–	1	Glucose metabolism is altered in cancers, including CRC (Fang and Fang, 2016).
hsa04144	Endocytosis	11.48	**0.00223191**	–	–	0.74809563	0.9971049	

(*Continued*)

TABLE 2 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04722	Neurotrophin signaling pathway	26.05	**0.00225063**	–	0.869732385	0.22082567	0.9303874	Neurotrophin signaling and related factors were found to clearly exert several biological and clinical features in CRC (Akil et al., 2016).
hsa04926	Relaxin signaling pathway	20.77	**0.002413722**	–	–	–	–	Relaxin signaling has a role in tumor cell growth and differentiation (Silvertown et al., 2003).
hsa04024	cAMP signaling pathway	14.57	**0.002417989**	0.37342574	–	–	–	Dysregulation cAMP signaling was implicated in many cancer types, including CRC (Löffler et al., 2008; Fajardo et al., 2014).
hsa04310	Wnt signaling pathway	17.72	**0.002418113**	–	0.068851256	0.3056234	1	Wnt signaling is a key player in many cancers, responsible for maintenance of cancer stem cells, metastasis and immune control (Zhan et al., 2017).
hsa05226	Gastric cancer	30.87	**0.00267019**	–	–	–	–	
hsa04392	Hippo signaling pathway - multiple species	17.24	**0.002915258**	–	–	–	–	Hippo signaling is involved in the control of intestinal stem cell proliferation and colorectal cancer development (Wierzbicki and Rybarczyk, 2015).
hsa04390	Hippo signaling pathway	16.23	**0.003135632**	–	–	–	–	Hippo signaling is involved in the control of intestinal stem cell proliferation and colorectal cancer development (Wierzbicki and Rybarczyk, 2015).
hsa00630	Glyoxylate and dicarboxylate metabolism	0	**0.003236188**	0.30407411	–	–	–	
hsa04110	Cell cycle	23.39	**0.003442925**	–	0.987280486	–	0.9898676	Dysregulation of the cell cycle is implicated in the biology of many cancers, including CRC (Hartwell and Kastan, 1994; Collins et al., 1997; Jarry et al., 2004).
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	13.42	**0.003808796**	–	–	–	–	
hsa05142	Chagas disease (American trypanosomiasis)	21.36	**0.003899445**	–	0.937326751	–	–	
hsa00410	beta-Alanine metabolism	3.23	**0.005120816**	**0.00196409**	–	–	1	
hsa04670	Leukocyte transendothelial migration	16.07	**0.005646255**	0.35563014	0.167771421	0.27631387	1	
hsa00620	Pyruvate metabolism	5.13	**0.00565919**	0.09639534	–	–	–	Glucose metabolism is altered in cancers, including CRC (Fang and Fang, 2016).
hsa04114	Oocyte meiosis	7.2	**0.006872183**	–	0.792716868	0.72884667	0.97175264	
hsa05215	Prostate cancer	51.55	**0.007778647**	–	0.598712628	0.32108408	1	

(*Continued*)

TABLE 2 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04210	Apoptosis	22.06	**0.007963488**	–	0.869732385	–	–	Abnormalities in apoptotic function contribute to both the pathogenesis of colorectal cancer and its resistance to chemotherapeutic drugs and radiotherapy (Watson, 2004).
hsa05140	Leishmaniasis	10.81	**0.008480037**	–	0.999999998	0.24636032	1	
hsa05222	Small cell lung cancer	27.96	**0.008933391**	–	0.120809416	0.45156074	1	
hsa05160	Hepatitis C	22.58	**0.010464676**	–	0.869732385	–	–	
hsa05031	Amphetamine addiction	13.24	**0.010805065**	–	0.107609007	–	–	
hsa04621	NOD-like receptor signaling pathway	6.74	**0.011387668**	–	0.999999998	0.5148187	0.9774175	NOD-like receptors are accepted as master regulators of inflammation and cancer (Saxena and Yeretssian, 2014).
hsa04914	Progesterone-mediated oocyte maturation	16.16	**0.011933047**	–	0.857467386	0.5950144	0.9922697	
hsa04923	Regulation of lipolysis in adipocytes	18.52	**0.011957867**	0.19643837	–	–	–	Adipocytes activate mitochondrial fatty acid oxidation and autophagy to promote tumor growth in colon cancer (Wen et al., 2017).
hsa04071	Sphingolipid signaling pathway	18.64	**0.012088886**	–	–	–	–	Sphingolipids have emerging roles in CRC (García-Barros et al., 2014).
hsa05016	Huntington disease	10.88	**0.013246653**	–	0.494422017	–	1	
hsa05030	Cocaine addiction	16.33	**0.014430369**	–	0.310528247	–	–	
hsa04270	Vascular smooth muscle contraction	12.4	**0.014703261**	**0.01450931**	**<0.001**	0.31157959	0.91536194	
hsa04915	Estrogen signaling pathway	22.06	**0.014973032**	–	–	–	–	
hsa04664	Fc epsilon RI signaling pathway	29.41	**0.016512816**	–	0.552502996	0.7568524	0.99502826	
hsa05211	Renal cell carcinoma	44.93	**0.017251888**	–	0.107609007	0.59724545	1	
hsa05202	Transcriptional misregulation in cancer	44.09	**0.017926078**	–	0.329766057	–	–	Core cancer pathway
hsa04913	Ovarian steroidogenesis	4.08	**0.021805547**	–	–	–	–	
hsa04620	Toll-like receptor signaling pathway	14.42	**0.023494048**	–	0.968714181	0.44691193	1	Toll-like receptor signaling pathway is being considered as a potential therapeutic target in colorectal cancer (Moradi-Marjaneh et al., 2018).
hsa04370	VEGF signaling pathway	33.9	**0.025228423**	–	0.768939947	0.53752804	1	Dysregulation of VEGF signaling is observed in numerous cancers, including CRC (Sun, 2012; Stacker and Achen, 2013).

(*Continued*)

TABLE 2 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04020	Calcium signaling pathway	9.57	**0.025565655**	0.33050764	0.057419238	0.3367621	0.9716838	Alterations of calcium signaling modulate tumor initiation, angiogenesis, progression and metastasis (Cui et al., 2017).
hsa05224	Breast cancer	31.29	**0.028494806**	–	–	–	–	
hsa04630	JAK-STAT signaling pathway	24.07	**0.029068433**	–	0.494422017	0.40343955	1	Jak-STAT signaling is involved in immune function and cell growth and has an important role in colorectal cancer (Slattery et al., 2013).
hsa04723	Retrograde endocannabinoid signaling	4.05	**0.029254258**	–	0.147248603	–	–	
hsa04622	RIG-I-like receptor signaling pathway	7.14	**0.030848585**	–	0.524757851	0.95792913	–	RIG-I-like receptors are important in immune signaling (Loo and Gale, 2011).
hsa04720	Long-term potentiation	22.39	**0.031734969**	–	0.899457922	0.7634754	0.9743132	
hsa04360	Axon guidance	14.36	**0.032363714**	0.14566283	**0.03397083**	0.31695387	0.98838806	
hsa04115	p53 signaling pathway	33.33	**0.033554867**	–	0.869732385	–	0.9952885	p53 signaling influences many key processes such as cell cycle arrest, apoptosis, and angiogenesis (Slattery et al., 2018).
hsa05131	Shigellosis	10.77	**0.033710491**	–	0.87420689	–	–	
hsa05203	Viral carcinogenesis	23.38	**0.036540488**	–	0.999999998	–	–	
hsa05416	Viral myocarditis	18.64	**0.038063956**	–	0.418726575	0.27175233	0.9940278	
hsa04666	Fc gamma R-mediated phagocytosis	20.88	**0.039418918**	–	0.141340043	0.32853782	–	
hsa00010	Glycolysis / Gluconeogenesis	1.47	**0.044512411**	0.35158586	–	–	1	Glucose metabolism is altered in cancers, including CRC (Fang and Fang, 2016).
hsa01212	Fatty acid metabolism	0	–	**<0.001**	–	–	1	
hsa01130	Biosynthesis of antibiotics	0	–	**<0.001**	–	–	–	
hsa04924	Renin secretion	7.69	0.050814742	**<0.001**	–	–	–	
hsa05414	Dilated cardiomyopathy	8.89	0.211547395	0.10754894	0.009508921	0.29030624	0.95637035	
hsa03320	PPAR signaling pathway	4.05	–	0.11340534	0.015730997	0.5118186	0.98862046	PPARδ acts as a tumor suppressor in colorectal cancer (You et al., 2015).
hsa01200	Carbon metabolism	0	–	**0.003293423**	–	–	–	
hsa01100	Metabolic pathways	0	–	**0.015541794**	–	–	–	Metabolic reprogramming has consequences at the cellular and molecular level with implications for cancer initiation and growth (Hagland et al., 2013)

"ID" indicates the Kyoto Encyclopedia of Genes and Genomes (KEGG) ID for the enriched pathway, whereas "Pathway" indicates the KEGG pathway name. "% CGC genes" indicates the percentage of Cancer Gene Census (CGC) genes in the pathway. The lowest Bonferroni-adjusted p value for pathfindR analysis is provided in "pathfindR," the false discovery rate (FDR)-adjusted p value for Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis is provided in "DAVID," the FDR-adjusted p value for Signaling Pathway Impact Analysis (SPIA) is presented in "SPIA," and the FDR-adjusted p values for Gene Set Enrichment Analysis (GSEA) and GSEAPreranked are presented in "GSEA" and "GSEAPreranked," respectively. Significant p values (i.e., adjusted p value < 0.05) are given in bold font. "-" indicates the pathway was not found to be enriched by the given tool. If a pathway is relevant to CRC, a brief description of its relevance is provided in "Brief Description."

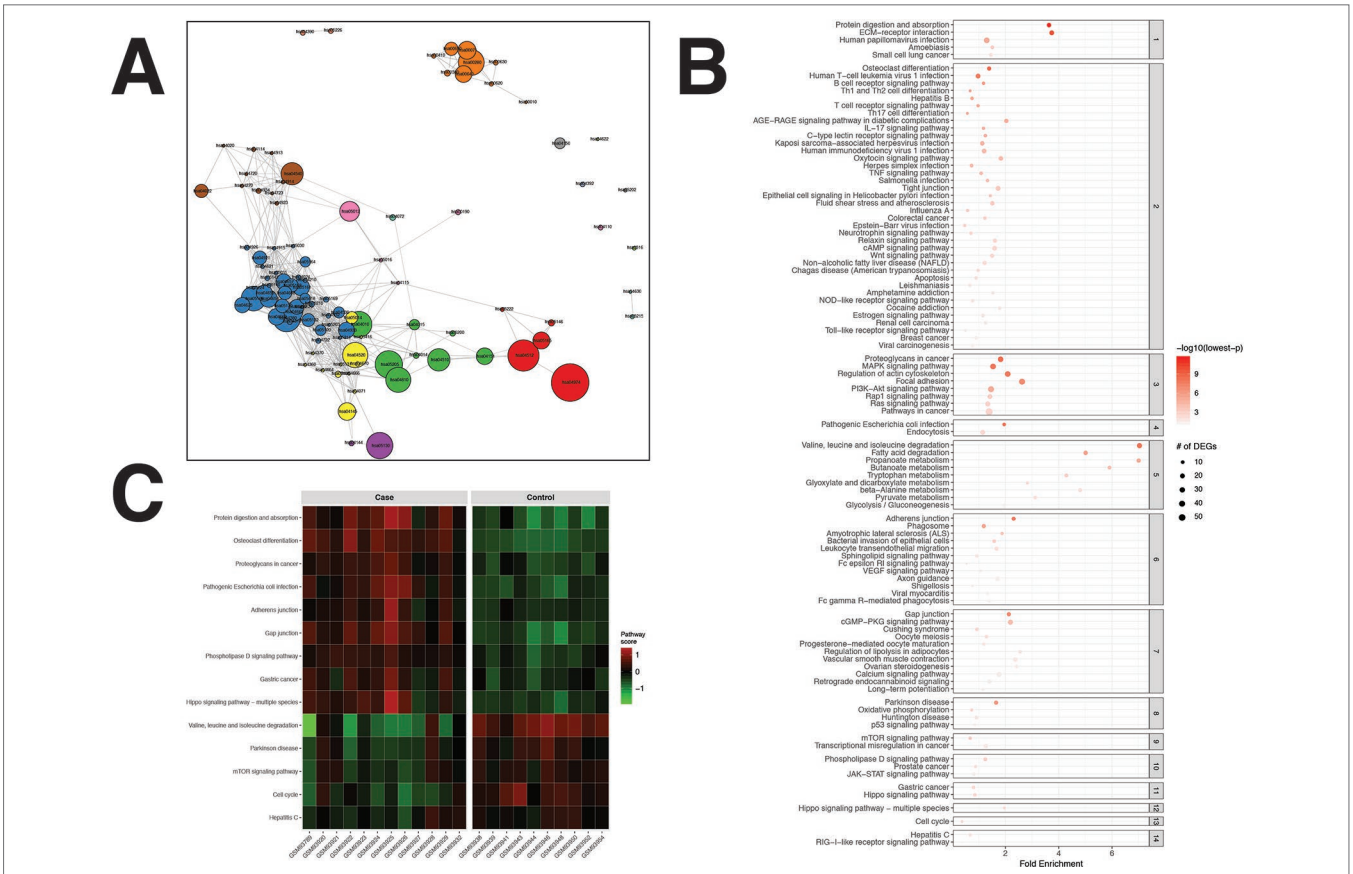


FIGURE 3 | pathfindR enrichment and clustering results on the colorectal cancer (CRC) dataset (lowest p ≤ 0.05). **(A)** Clustering graph, each color displaying the clusters obtained for CRC. Each node is an enriched pathway. The size of a node corresponds to its –log(lowest_p). The thickness of the edges between nodes corresponds to the kappa statistic between the two terms. **(B)** Bubble chart of enrichment results grouped by clusters (labeled on the right-hand side of each panel). The x axis corresponds to fold enrichment values, while the y axis indicates the enriched pathways. The size of the bubble indicates the number of differentially expressed genes (DEGs) in the given pathway. The color indicates the –log10(lowest-p) value; the more it shifts to red, the more significantly the pathway is enriched. **(C)** Heat map of pathway scores per subject. The x axis indicates subjects, whereas the y axis indicates representative pathways. Color scale for the pathway score is provided in the right-hand legend.

percentages of CGC genes of the enriched pathways of pathfindR were again the highest among all tools for all significance cutoff values (18.73%, 18.37%, and 17.93% for 0.05, 0.1, and 0.25, respectively).

Permutation Assessment

To assess the number of pathways identified to be enriched by pathfindR, we performed analyses using actual and permuted data of different sizes. Comparison of the distributions of actual vs. permuted data is presented in **Figure 6**. Wilcoxon rank sum tests revealed that the distributions of the numbers of enriched pathways obtained using actual and permuted input data were significantly different (all $p < 0.001$). The median number of enriched pathways was lower for permuted data in each case.

It was observed that the ratio of the median number of pathways (permuted/actual) tended to increase as the number of input genes increased. This is most likely because as the input size gets larger, there is higher chance in finding highly connected subnetworks that in turn leads to identifying a higher number of enriched pathways.

Assessment of the Effect of DEGs Without Any Interactions on Enrichment Results

To gain further support for our proposal that directly performing enrichment analysis on a list of genes is not completely informative because this ignores the interaction information, we performed ORA (as implemented in pathfindR) on (i) all of the DEG lists (RA, CRC, and PCa) and (ii) the filtered list of DEGs for the same datasets so that they only contain DEGs found in the Biogrid PIN. This allowed us to assess any effect of eliminating DEGs with no interactions on the enrichment results.

The numbers of DEGs found in the Biogrid PIN for each dataset was as follows: RA—481 (out of 572 total), CRC—989 (out of 1,356) and PCa—900 (out of 1,240). The ORA results are presented in **Supplementary Data Sheet 3**. The elimination of DEGs without any interaction clearly affected numbers of significantly enriched (FDR < 0.05) KEGG pathways (**Supplementary Table 2**). For the RA dataset, no significantly enriched pathways were found using all DEGs, whereas elimination of non-interacting DEGs resulted in one significant pathway. For CRC and PCa, using only DEGs found

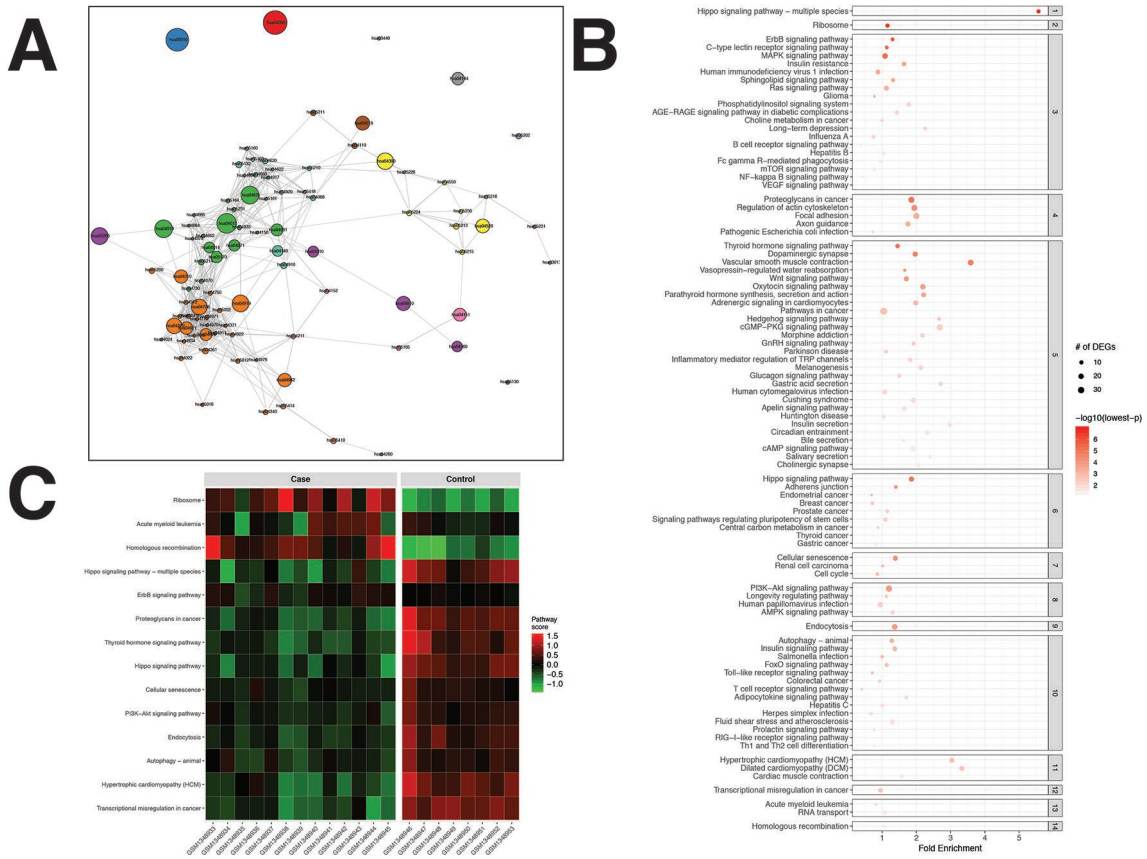


FIGURE 4 | pathfindR enrichment and clustering results on the prostate cancer (PCa) dataset (lowest p ≤ 0.05). **(A)** Clustering graph, each color displaying the clusters obtained for PCa. Each node is an enriched pathway. The size of a node corresponds to its –log(lowest_p). The thickness of the edges between nodes corresponds to the kappa statistic between the two terms. **(B)** Bubble chart of enrichment results grouped by clusters (labeled on the right-hand side of each panel). The x axis corresponds to fold enrichment values, while the y axis indicates the enriched pathways. The size of the bubble indicates the number of differentially expressed genes (DEGs) in the given pathway. The color indicates the –log10(lowest-p) value; the more it shifts to red, the more significantly the pathway is enriched. **(C)** Heat map of pathway scores per subject. The x axis indicates subjects, whereas the y axis indicates representative pathways. Color scale for the pathway score is provided in the right-hand legend.

in the PIN, the number of significantly enriched pathways were doubled compared to using all of the genes without taking into account any interaction information. We would like to note that these results partly explain why taking interaction information into account results in enhanced enrichment results.

Assessment of the Effect of PINs on Enrichment Results

To assess any effect of the choice of PIN on pathfindR results, we first compared the default PINs in terms of the interactions they contain. The number of interactions in the PINs were as follows: 289,417 interactions in Biogrid, 79,741 interactions in GeneMania, 121,007 interactions in IntAct, and 53,047 interactions in KEGG. The numbers of common interactions between any pair of PINs and the overlap percentages of the interactions are presented in **Supplementary Table 3**. The results show that there is very little overlap between the PINs. Despite the fact that Biogrid has more than double the interactions of IntAct and 3 times the interactions of GeneMania, it remarkably does not contain half of the interactions they contain, implying this lack of overlap between PINs may affect pathfindR results.

We then proceeded with analyzing any effect of the choice of PIN on active-subnetwork-oriented pathway enrichment analysis. Venn diagrams comparing enrichment results obtained through pathfindR analyses with all available PINs are presented in **Supplementary Figure 1**. This comparison revealed that there was no compelling overlap among the enriched pathways obtained by using different PINs. Overall, using Biogrid and KEGG resulted in the highest number of significantly enriched pathways for all datasets.

As described in Materials and Methods, the results presented in this subsection were obtained using greedy search with search depth of 1 and maximum depth of 1, which results in multiple subnetworks structured as local subnetworks. Although it is not fully dependent on it, this method requires direct interactions between input genes. In the extreme case where there is no direct connection between any pair of two input genes, it is impossible to get any multi-node subnetworks with this method. Therefore, in order to gain a better understanding of the lack of overlap between the enrichment results

TABLE 3 | Pathway analysis results for the prostate cancer (PCa) dataset (adjusted $p < 0.05$).

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04392	Hippo signaling pathway - multiple species	17.24	**<0.001**	–	–	–	–	The hippo pathway effector YAP regulates motility, invasion, and castration-resistant growth of prostate cancer cells (Zhang et al., 2015).
hsa03010	Ribosome	1.96	**<0.001**	–	–	0.425191	1	Certain ribosomal proteins are altered and may serve as putative biomarkers for prostate cancer (Arthurs et al., 2017).
hsa04012	ErbB signaling pathway	40	**<0.001**	–	0.637063484	–	–	There are interactions among the ErbB receptor network, its downstream pathways, and androgen receptor signaling (El Sheikh et al., 2003).
hsa04625	C-type lectin receptor signaling pathway	27.88	**<0.001**	–	–	–	–	C-type lectin receptors are emerging orchestrators of sterile inflammation and represent potential therapeutic targets in many cancers, including PCa (Chiffoleau, 2018). C-type lectins were shown to facilitate tumor metastasis (Ding et al., 2017).
hsa04010	MAPK signaling pathway	17.97	**<0.001**	–	0.282733912	–	0.90003437	MAPK signaling pathways act through their effects on apoptosis, survival, metastatic potential, and androgen-independent growth in prostate cancer (Rodríguez-Berriguete et al., 2012).
hsa05205	Proteoglycans in cancer	27.86	**<0.001**	**0.02399376**	–	–	–	Proteoglycans play roles in modulating cancer progression, invasion and metastasis (Iozzo and Sanderson, 2011).
hsa04919	Thyroid hormone signaling pathway	32.76	**<0.001**	0.5109672	–	–	–	
hsa04390	Hippo signaling pathway	16.23	**<0.001**	0.10679672	–	–	–	The hippo pathway effector YAP regulates motility, invasion, and castration-resistant growth of prostate cancer cells (Zhang et al., 2015).
hsa04728	Dopaminergic synapse	11.45	**<0.001**	0.06897641	**0.034643839**	–	–	
hsa04270	Vascular smooth muscle contraction	12.4	**<0.001**	**<0.001**	**<0.001**	–	0.9954409	
hsa04810	Regulation of actin cytoskeleton	15.02	**<0.001**	**0.02727598**	**0.033699531**	–	–	Dysregulated in cancer cell migration and invasion (Yamaguchi and Condeelis, 2007).
hsa04218	Cellular senescence	25.63	**<0.001**	–	–	–	–	Cellular senescence may play a role in treatment resistance in PCa (Blute et al., 2017).

(*Continued*)

TABLE 3 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04520	Adherens junction	31.94	**<0.001**	–	–	–	–	Dysregulation of the adherens junction system has particular implications in transformation and tumor invasion (Knights et al., 2012).
hsa04962	Vasopressin-regulated water reabsorption	13.64	**<0.001**	–	0.655412336	–	–	
hsa04310	Wnt signaling pathway	17.72	**<0.001**	0.14714863	0.174150166	–	–	Wnt signaling is implicated in PCa biology (Murillo-Garzon and Kypta, 2017).
hsa04151	PI3K-Akt signaling pathway	21.47	**<0.001**	–	–	–	–	Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion (Shukla et al., 2007).
hsa04921	Oxytocin signaling pathway	13.16	**<0.001**	0.09939094	–	–	–	Oxytocin signaling has a role in prostate cancer metastasis (Zhong et al., 2010).
hsa04144	Endocytosis	11.48	**<0.001**	0.14183304	–	–	–	Defective vesicular trafficking of growth factor receptors, as well as unbalanced recycling of integrin- and cadherin-based adhesion complexes, has emerged as a multifaceted hallmark of malignant cells (Mosesson et al., 2008).
hsa04928	Parathyroid hormone synthesis, secretion and action	20.75	**<0.001**	–	–	–	–	
hsa04931	Insulin resistance	14.81	**<0.001**	0.44248049	–	–	–	Men in the highest tertile of insulin resistance (IR) had an increased risk of prostate cancer, indicating a potential pathogenetic link of IR with prostate cancer (Hsing et al., 2003).
hsa05170	Human immunodeficiency virus 1 infection	16.98	**<0.001**	–	–	–	–	
hsa04071	Sphingolipid signaling pathway	18.64	**<0.001**	–	–	–	–	Sphingolipids are modulators of cancer cell death and represent potential therapeutic targets (Segui et al., 2006; Shaw et al., 2018).
hsa04510	Focal adhesion	19.1	**<0.001**	**0.01182864**	**0.003797795**	–	–	Cancer cells exhibit highly altered focal adhesion dynamics (Maziveyi and Alahari, 2017).
hsa04014	Ras signaling pathway	18.1	**<0.001**	–	–	–	–	Ras signaling plays an important role in prostate cancer progression and is a possibly mediator of hormone resistance (Weber and Gioeli, 2004; Whitaker and Neal, 2010).

(*Continued*)

TABLE 3 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04140	Autophagy - animal	17.19	**<0.001**	–	0.91466497	0.7367432	–	Autophagy is a modulator of PCa biology and is a therapeutic target (Farrow et al., 2014).
hsa04360	Axon guidance	14.36	**<0.001**	0.36615434	0.174150166	–	–	
hsa04910	Insulin signaling pathway	18.98	**<0.001**	–	0.592610905	–	–	Insulin signaling has crucial roles in cell proliferation and death. Insulin receptors were detected on primary human prostate cancers (Cox et al., 2009; Bertuzzi et al., 2016).
hsa05132	Salmonella infection	8.14	**0.001024926**	–	0.884388639	–	–	
hsa04261	Adrenergic signaling in cardiomyocytes	11.81	**0.001251519**	0.1359051	–	–	–	
hsa05213	Endometrial cancer	60.34	**0.001571998**	–	0.889535144	0.9776995	0.9350631	
hsa05211	Renal cell carcinoma	44.93	**0.001704596**	–	0.958690885	–	–	
hsa05200	Pathways in cancer	26.81	**0.001864931**	0.44205232	0.592610905	–	–	"Meta"-pathway of cancer pathways.
hsa05214	Glioma	44	**0.00191144**	–	0.678606672	–	–	
hsa04110	Cell cycle	23.39	**0.00200072**	–	0.53576482	0.73860705	–	Dysregulation of the cell cycle is implicated in the biology of many cancers, including PCa (Hartwell and Kastan, 1994; Collins et al., 1997; Balk and Knudsen, 2008).
hsa05410	Hypertrophic cardiomyopathy (HCM)	6.02	**0.002088682**	**0.01508581**	–	–	0.94539815	
hsa05202	Transcriptional misregulation in cancer	44.09	**0.002227785**	–	0.909985754	–	–	Core cancer pathway.
hsa04068	FoxO signaling pathway	29.55	**0.00256445**	–	–	–	–	FOXO signaling is implicated and considered as a therapeutic target in many cancers, including PCa (Farhan et al., 2017).
hsa04620	Toll-like receptor signaling pathway	14.42	**0.002757387**	–	0.999737262	1	–	TLRs may serve as a double-edged sword in prostate cancer tumorigenesis by promoting malignant transformation of epithelial cells and tumor growth, or on the contrary, inducing apoptosis, and inhibiting tumor progression (Zhao et al., 2014).
hsa05414	Dilated cardiomyopathy (DCM)	8.89	**0.002884316**	**0.00445895**	**0.00145624**	–	0.9310661	
hsa05224	Breast cancer	31.29	**0.002996465**	–	–	–	–	
hsa04340	Hedgehog signaling pathway	21.28	**0.003701704**	–	0.603911642	–	–	Hedgehog signaling plays an important role in the development and progression of PCa (Gonnissen et al., 2013).
hsa05215	Prostate cancer	51.55	**0.004678127**	–	0.637063484	–	–	The pathway of the disease.

(*Continued*)

TABLE 3 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04211	Longevity regulating pathway	25.84	**0.004704898**	–	–	–	–	
hsa04022	cGMP-PKG signaling pathway	11.04	**0.004947514**	**0.00142051**	–	–	–	cGMP-PKG signaling inhibits cell proliferation and induces apoptosis (Fajardo et al., 2014).
hsa05032	Morphine addiction	4.4	**0.005138281**	0.37736294	0.174150166	–	–	
hsa04550	Signaling pathways regulating pluripotency of stem cells	31.65	**0.00540706**	–	–	–	–	
hsa04912	GnRH signaling pathway	19.35	**0.005600378**	0.37736294	0.340227111	–	0.9284794	GnRH signaling has roles in cancer cell proliferation and metastasis in many cancers, including PCa (Gründker and Emons, 2017).
hsa05165	Human papillomavirus infection	19.09	**0.005787239**	–	–	–	–	HPV infection is associated with increasing risk of PCa, indicating a potential pathogenetic link between HPV and prostate cancer (Yin et al., 2017).
hsa05012	Parkinson disease	6.34	**0.005882045**	–	0.895565575	–	–	
hsa04070	Phosphatidylinositol signaling system	6.06	**0.007170858**	0.47255633	0.592215095	–	–	Deregulation PI3 kinase signaling is implicated in prostate carcinogenesis (Elfiky and Jiang, 2013).
hsa04750	Inflammatory mediator regulation of TRP channels	10.1	**0.007170858**	0.47255633	–	–	–	TRP channels have emerged as key proteins in central mechanisms of the carcinogenesis such as cell proliferation, apoptosis and migration (Gkika and Prevarskaya, 2011).
hsa04933	AGE-RAGE signaling pathway in diabetic complications	31	**0.007170858**	–	–	–	–	
hsa05231	Choline metabolism in cancer	27.27	**0.007170858**	–	–	–	–	Core cancer pathway. Choline metabolites can be used as potential prognostic biomarkers for the management of prostate cancer patients (Awwad et al., 2012).
hsa04730	Long-term depression	23.33	**0.007411265**	0.29433246	0.228920589	–	–	
hsa04152	AMPK signaling pathway	15.83	**0.007412407**	–	–	–	–	First identified as a master regulator of metabolism, AMPK may have numerous roles beyond metabolism. AMPK signaling can have context-dependent effects in prostate cancer (Khan and Frigo, 2017).
hsa05210	Colorectal cancer	51.16	0.007572536	–	0.53576482	–	0.8859366	
hsa04660	T cell receptor signaling pathway	31.68	0.007759806	–	0.999737262	–	–	T-cell receptor signaling modulates control of anti-cancer immunity (Cronin and Penninger, 2007).
hsa04916	Melanogenesis	20.79	0.007759806	0.23117007	0.191012563	–	–	

(*Continued*)

TABLE 3 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04922	Glucagon signaling pathway	11.65	**0.008383558**	–	–	–	–	
hsa04971	Gastric acid secretion	8	**0.008625239**	0.17201492	0.174150166	–	–	
hsa05164	Influenza A	15.79	**0.009418672**	–	0.871606688	–	–	
hsa05230	Central carbon metabolism in cancer	49.23	**0.00943342**	–	–	–	–	Core cancer pathway.
hsa05163	Human cytomegalovirus infection	24	**0.010897057**	–	–	–	–	
hsa04920	Adipocytokine signaling pathway	18.84	**0.011289828**	–	0.573213367	–	–	Adipocytokines are implicated in many cancers, including PCa (Housa et al., 2006).
hsa05130	Pathogenic Escherichia coli infection	10.91	**0.012638019**	–	0.914414969	–	–	
hsa05160	Hepatitis C	22.58	**0.012701709**	–	0.952731561	–	–	
hsa05168	Herpes simplex infection	10.81	**0.012818879**	–	0.999737262	–	–	
hsa04934	Cushing syndrome	22.73	**0.012968007**	–	–	–	–	
hsa04662	B cell receptor signaling pathway	42.25	**0.015323796**	–	0.871606688	–	–	
hsa05418	Fluid shear stress and atherosclerosis	18.71	**0.016016719**	–	–	–	–	
hsa05216	Thyroid cancer	70.27	**0.016672034**	–	0.77019655	–	–	
hsa05221	Acute myeloid leukemia	50	**0.018273818**	–	0.916809245	0.9737256	0.9253648	
hsa04371	Apelin signaling pathway	13.14	**0.019388476**	–	–	–	–	Various apelin peptides can stimulate tumor growth and proliferation of many types of cancer cells, including PCa (Wysocka et al., 2018).
hsa05016	Huntington disease	10.88	**0.019575207**	–	0.887106943	1	0.9790702	
hsa04911	Insulin secretion	11.76	**0.021091491**	**0.02334547**	–	–	–	
hsa04917	Prolactin signaling pathway	31.43	**0.021797194**	–	–	–	–	Prolactin signalling promotes prostate tumorigenesis and may be targeted for therapy (Goffin et al., 2011; Sackmann-Sala and Goffin, 2015).
hsa03440	Homologous recombination	24.39	**0.027879334**	–	–	0.82518643	0.88681024	Homologous recombination offers a model for novel DNA repair targets and therapies in PCa (Bristow et al., 2007).
hsa04713	Circadian entrainment	10.31	**0.028299959**	0.11376372	–	–	–	
hsa03013	RNA transport	5.45	**0.029590861**	–	0.887106943	–	–	Many common and specialized mRNA export factors are dysregulated in cancer (Siddiqui and Borden, 2012).
hsa04260	Cardiac muscle contraction	6.41	**0.030119546**	–	–	–	0.92371947	
hsa05161	Hepatitis B	31.29	**0.031074222**	–	–	–	–	
hsa04666	Fc gamma R-mediated phagocytosis	20.88	**0.032172546**	–	0.838868067	–	–	
hsa04976	Bile secretion	4.23	**0.032183139**	–	0.77019655	–	–	

(*Continued*)

TABLE 3 | Continued

ID	Pathway	% CGC genes	pathfindR	DAVID	SPIA	GSEA	GSEAPreranked	Brief Description
hsa04024	cAMP signaling pathway	14.57	**0.035154932**	0.12151133	–	–	–	Dysregulation cAMP signaling was implicated in many cancer types, including PCa (Fajardo et al., 2014).
hsa05226	Gastric cancer	30.87	**0.035466235**	–	–	–	–	
hsa04622	RIG-I-like receptor signaling pathway	7.14	**0.036168176**	–	0.678606672	0.998692	1	
hsa04150	mTOR signaling pathway	16.45	**0.03639603**	–	0.608898009	0.97279966	0.8907857	mTOR signaling is implicated in prostate cancer progression and androgen deprivation therapy resistance (Edlind and Hsieh, 2014).
hsa04064	NF-kappa B signaling pathway	17.89	**0.036565869**	–	0.999737262	–	–	The NF-kappa B signaling pathway controls the progression of Pca (Jin et al., 2008).
hsa04970	Salivary secretion	6.67	**0.038144831**	0.35044831	0.228920589	–	–	
hsa04658	Th1 and Th2 cell differentiation	19.57	**0.040720473**	–	–	–	–	T helper cells are important in cancer immunity (Knutson and Disis, 2005).
hsa04370	VEGF signaling pathway	33.9	**0.043130708**	–	0.889535144	–	–	Angiogenesis has been shown to play an important role in tumorigenesis, proliferation and metastasis in PCa. Various promising agents that target VEGF signaling have been tested (Aragon-Ching and Dahut, 2009).
hsa04725	Cholinergic synapse	18.75	**0.04793374**	0.13876451	0.129551973	–	–	
hsa00120	Primary bile acid biosynthesis	0	–	–	–	0.78211117	**<0.001**	

"ID" indicates the Kyoto Encyclopedia of Genes and Genomes (KEGG) ID for the enriched pathway, whereas "Pathway" indicates the KEGG pathway name. "% CGC genes" indicates the percentage of Cancer Gene Census (CGC) genes in the pathway. The lowest Bonferroni-adjusted p value for pathfindR analysis is provided in "pathfindR," the false discovery rate (FDR)-adjusted p value for Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis is provided in "DAVID," the FDR-adjusted p value for Signaling Pathway Impact Analysis (SPIA) is presented in "SPIA," and the FDR-adjusted p values for Gene Set Enrichment Analysis (GSEA) and GSEAPreranked are presented in "GSEA" and "GSEAPreranked," respectively. Significant p values (i.e., adjusted p value < 0.05) are given in bold font. "-" indicates the pathway was not found to be enriched by the given tool. If a pathway is relevant to PCa, a brief description of its relevance is provided in "Brief Description."

presented above, we analyzed the numbers of direct interactions of input genes in each PIN. These results are presented as Venn diagrams in **Supplementary Figure 2**. It is striking that there are only nine common interactions of RA DEGs in all PINs (although there are 54 common interactions in PINs except KEGG). The findings are similar for the CRC and PCa datasets: there are 11 common CRC DEG interactions in all PINs (81 in PINs except KEGG), and 5 PCa DEG interactions (56 in PINs except KEGG).

In case of utilizing KEGG PIN and KEGG pathways, the same interactions for both subnetwork interaction and enrichment analysis are considered. This approach does not introduce any extra information to the analysis, and it is clear that interacting gene groups in the KEGG PIN will be enriched in KEGG pathways. This explains the high number of pathways obtained using the KEGG PIN. Moreover, it is known that pathways in pathway databases may be strongly biased by some classes of genes or phenotypes that are popular targets, such as cancer signaling (Liu et al., 2017a). Therefore, the PIN obtained through KEGG pathway interactions are biased. Biogrid has the highest coverage for direct interactions among DEGs as seen in **Supplementary Figure 2**. It is unbiased in terms of phenotypes, and using Biogrid to extract KEGG pathways combines the two sources of information.

Considering all of the above-mentioned findings, we conclude that utilizing the Biogrid PIN can provide the researcher with the most extensive enrichment results.

DISCUSSION

PathfindR is an R package that enables active subnetwork-oriented pathway analysis, complementing the gene-phenotype associations identified through differential expression/methylation analysis.

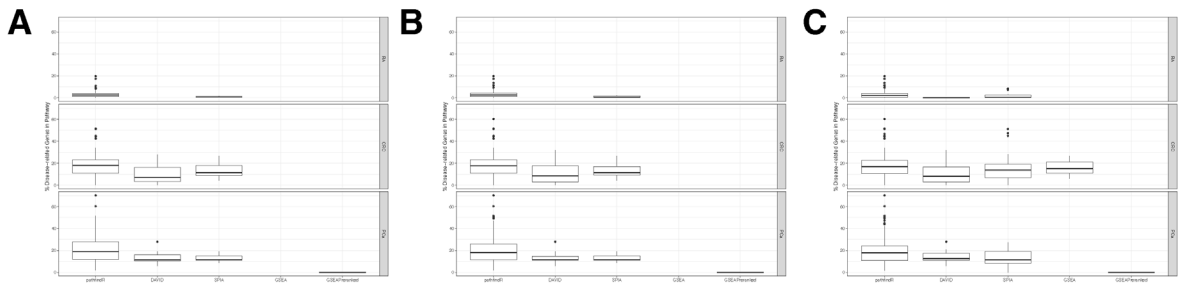
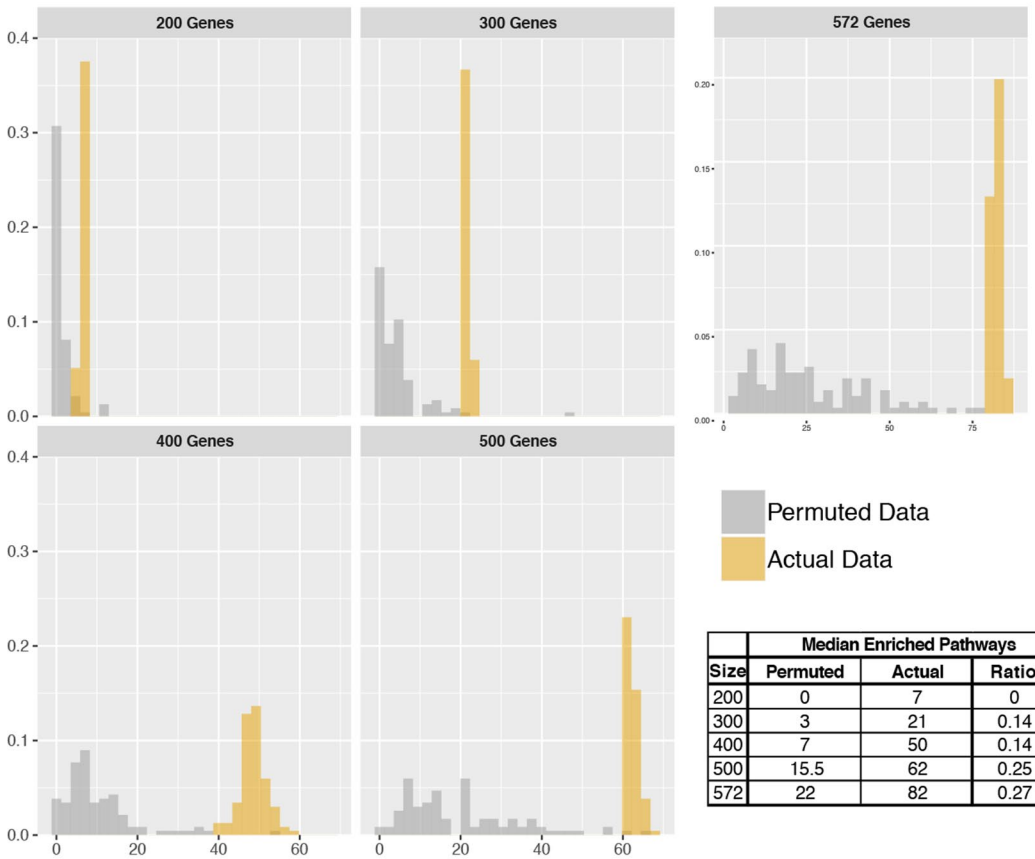


FIGURE 5 | Distributions of disease-associated genes in the enriched pathways. Boxplots displaying the distributions of the percentages of disease-related genes in the pathways found to be enriched by pathfindR, Database for Annotation, Visualization and Integrated Discovery (DAVID), Signaling Pathway Impact Analysis (SPIA), Gene Set Enrichment Analysis (GSEA), and GSEAPreranked in the datasets rheumatoid arthritis (RA), colorectal cancer (CRC), and prostate cancer (PCa). No boxplot for a tool in a particular dataset indicates that the given tool did not identify any enriched pathways in the given dataset. **(A)** Boxplots for all the results filtered for adjusted-p ≤ 0.05. **(B)** Boxplots for all the results filtered for adjusted-p ≤ 0.1. **(C)** Boxplots for all the results filtered for adjusted-p ≤ 0.25.



Size	Median Enriched Pathways			Wilcoxon p value
	Permuted	Actual	Ratio	
200	0	7	0	8.4e-31
300	3	21	0.14	4.2e-33
400	7	50	0.14	2.69e-33
500	15.5	62	0.25	1.04e-33
572	22	82	0.27	1.04e-33

FIGURE 6 | Distributions of the number of enriched pathways for actual vs. permuted data. Histograms displaying the distributions of the number of enriched pathways for actual and permuted data for input sizes of 200, 300, 400, 500, and 572. The x axes correspond to the number of enriched pathways, and the y axes correspond to relative frequencies. On the right bottom, a table summarizing the results is provided.

In most gene set enrichment approaches, relational information captured in the graph structure of a PIN is overlooked. Hence, during these analyses, genes in the network neighborhood of significant genes are not taken into account. The approach we considered for exploiting interaction information to enhance pathway enrichment analysis was active subnetwork search. In a nutshell, active subnetwork search enables inclusion of genes that are not significant genes themselves but connect significant genes. This results in the identification of phenotype-associated connected significant subnetworks. Initially identifying active subnetworks in a list of significant genes and then performing pathway enrichment analysis of these active subnetworks efficiently

exploits interaction information between the genes. This, in turn, helps uncover relevant phenotype-related mechanisms underlying the disease, as demonstrated in the example applications.

Through pathfindR, numerous relevant pathways were identified in each example. The literature-supported disease-related pathways mostly ranked higher in the pathfindR results. The majority of additional pathways identified through pathfindR were relevant to the pathogenesis of the diseases under study, as supported by literature. A separate confirmation of disease-relatedness was provided by analysis of the distributions of the percentage of disease genes in the identified pathways. This analysis revealed that pathfindR pathways contained the highest median percentages of disease-related genes in each dataset regardless of significance cutoff value, implying that the pathways identified by pathfindR are indeed associated with the given disease. Together, these two assessments of disease-relatedness of pathways indicate that pathfindR produces pathway enrichment results at least as relevant as the other tools widely used for enrichment analysis.

We propose that pathfindR performed better than the analyzed pathway analysis tools because, for enrichment analysis, it included disease-related genes that were not in the DEG list but that were known to interact with the DEGs, which most enrichment tools disregard. By performing enrichment analyses on distinct sets of interacting genes (i.e., active subnetworks), pathfindR also eliminated "false positive" genes that lacked any strong interaction. The above findings indicate that incorporating interaction information prior to enrichment analysis results in better identification of disease-related mechanisms.

This package extends the use of the active-subnetwork-oriented pathway analysis approach to omics data. Additionally, it provides numerous improvements and useful new features. The package provides three active subnetwork search algorithms. The researcher is therefore able to choose between the different algorithms to obtain the optimal results. For the greedy and simulated annealing active subnetwork search algorithms, the search and enrichment processes are executed several times. By summarizing results over the iterations and identifying consistently enriched pathways, the stochasticity of these algorithms is overcome. Additionally, the researcher is able to choose from several built-in PINs and can use their own custom PIN by providing the path to the SIF file. The researcher is also able to choose from numerous built-in gene sets, listed above, and can also provide a custom gene set resource. pathfindR also allows for clustering of related pathways. This allows for combining relevant pathways together, uncovering coherent "meta-pathways" and reducing complexity for easier interpretation of findings. This clustering functionality also aids in eliminating falsely enriched pathways that are initially found because of their similarity to the actual pathway of interest. The package also allows for scoring of pathways in individual subjects, denoting the pathway activity. Finally, pathfindR is built as a stand-alone package, but it can easily be integrated with other tools, such as differential expression/methylation analysis tools, for building fully automated pipelines.

To the best of our knowledge, pathfindR is the first and, so far, the only R package for active-subnetwork-oriented pathway enrichment analysis. It also offers functionality for pathway clustering, scoring, and visualization. All features in pathfindR work together to enable identification and further investigation of dysregulated pathways that potentially reflect the underlying pathological mechanisms. We hope that this approach will allow researchers to better answer their research questions and discover mechanisms underlying the phenotype being studied.

AUTHOR CONTRIBUTIONS

OS, OO, and EU conceived the pathway analysis approach. OO and EU implemented the R package. EU performed the analyses presented in this article. OUS, OO, and EU interpreted the results. All authors were involved in the writing of the manuscript and read and approved the version being submitted.

ACKNOWLEDGMENTS

This manuscript has been initially released as a pre-print at bioRxiv (Ulgen et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00858/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Venn diagram of enrichment results obtained through pathfindR analyses with all available PINs.

SUPPLEMENTARY FIGURE 2 | Venn diagram of the numbers of direct interactions of input genes in each PIN.

SUPPLEMENTARY DATASHEET 1 | The results of differential expression analyses for RA, CRC and PCa, prior to filtering (differential expression statistics for all probes) and after filtering (lists of DEGs).

SUPPLEMENTARY DATASHEET 2 | The unfiltered results of enrichment analyses using all the different methods on each of the datasets.

SUPPLEMENTARY DATASHEET 3 | ORA results using all DEGs and using only DEGs found in the BioGRID PIN for each dataset.

REFERENCES

Akil, H., Perraud, A., Jauberteau, M.-O., and Mathonnet, M. (2016). Tropomyosin-related kinase B/brain derived-neurotrophic factor signaling pathway as a potential therapeutic target for colorectal cancer. *World J. Gastroenterol.* 22 (2), 490–500. doi: 10.3748/wjg.v22.i2.490

Amgalan, B., and Lee, H. (2014). WMAXC: a weighted maximum clique method for identifying condition-specific sub-network. *PLoS One* 9 (8), e104993. doi: 10.1371/journal.pone.0104993

Antanaviciute, I., Mikalayeva, V., Cesleviciene, I., Milasiute, G., Skeberdis, V. A., and Bordel, S. (2017). Transcriptional hallmarks of cancer cell lines reveal an emerging role of branched chain amino acid catabolism. *Sci. Rep.* 7 (1), 7820. doi: 10.1038/s41598-017-08329-8

Aragon-Ching, J. B., and Dahut, W. L. (2009). VEGF inhibitors and prostate cancer therapy. *Curr. Mol. Pharmacol.* 2 (2), 161–168. doi: 10.2174/1874467210902020161

Arthurs, C., Murtaza, B. N., Thomson, C., Dickens, K., Henrique, R., Patel, H. R. H., et al. (2017). Expression of ribosomal proteins in normal and cancerous human

prostate tissue. *PLoS One* 12 (10), e0186047. doi: 10.1371/journal.pone.0186047

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consortium. Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Awwad, H. M., Geisel, J., and Obeid, R. (2012). The role of choline in prostate cancer. *Clin. Biochem.* 45 (18), 1548–1553. doi: 10.1016/j.clinbiochem.2012.08.012

Backes, C., Rurainski, A., Klau, G. W., Muller, O., Stockel, D., Gerasch, A., et al. (2012). An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Res.* 40 (6), e43. doi: 10.1093/nar/gkr1227

Bakir-Gungor, B., Baykan, B., Ugur Iseri, S., Tuncer, F. N., and Sezerman, O. U. (2013). Identifying SNP targeted pathways in partial epilepsies with genome-wide association study data. *Epilepsy Res.* 105 (1–2), 92–102. doi: 10.1016/j.eplepsyres.2013.02.008

Bakir-Gungor, B., Egemen, E., and Sezerman, O. U. (2014). PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data. *Bioinformatics* 30 (9), 1287–1289. doi: 10.1093/bioinformatics/btt743

Bakir-Gungor, B., Remmers, E. F., Meguro, A., Mizuki, N., Kastner, D. L., Gul, A., et al. (2015). Identification of possible pathogenic pathways in Behcet's disease using genome-wide association study data from two different populations. *Eur. J. Hum. Genet.* 23 (5), 678–687. doi: 10.1038/ejhg.2014.158

Bakir-Gungor, B., and Sezerman, O. U. (2013). The identification of pathway markers in intracranial aneurysm using genome-wide association data from two different populations. *PLoS One* 8 (3), e57022. doi: 10.1371/journal.pone.0057022

Balk, S. P., and Knudsen, K. E. (2008). AR, the cell cycle, and prostate cancer. *Nucl. Recept Signal* 6. doi: 10.1621/nrs.06001

Barthel, C., Yeremenko, N., Jacobs, R., Schmidt, R. E., Bernateck, M., Zeidler, H., et al. (2009). Nerve growth factor and receptor expression in rheumatoid arthritis and spondyloarthritis. *Arthritis Res. Ther.* 11 (3), R82. doi: 10.1186/ar2716

Beisser, D., Brunkhorst, S., Dandekar, T., Klau, G. W., Dittrich, M. T., and Muller, T. (2012). Robustness and accuracy of functional modules in integrated network analysis. *Bioinformatics* 28 (14), 1887–1894. doi: 10.1093/bioinformatics/bts265

Berridge, M. J. (2016). The inositol trisphosphate/calcium signaling pathway in health and disease. *Physiol. Rev.* 96 (4), 1261–1296. doi: 10.1152/physrev.00006.2016

Bertuzzi, A., Conte, F., Mingrone, G., Papa, F., Salinari, S., and Sinisgalli, C. (2016). Insulin signaling in insulin resistance states and cancer: a modeling analysis. *PLoS One* 11 (5), e0154415. doi: 10.1371/journal.pone.0154415

Bigelow, K., and Nguyen, T. A. (2014). Increase of gap junction activities in SW480 human colorectal cancer cells. *BMC Cancer* 14, 502. doi: 10.1186/1471-2407-14-502

Blute, M. L., Jr., Damaschke, N., Wagner, J., Yang, B., Gleave, M., Fazli, L., et al. (2017). Persistence of senescent prostate cancer cells following prolonged neoadjuvant androgen deprivation therapy. *PloS One* 12 (2), e0172048–e0172048. doi: 10.1371/journal.pone.0172048

Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., et al. (2014). Colonization of the human gut by E. coli and colorectal cancer risk. *Clin. Cancer Res.* 20 (4), 859–867. doi: 10.1158/1078-0432.CCR-13-1343

Breitling, R., Amtmann, A., and Herzyk, P. (2004). Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinformatics* 5, 100. doi: 10.1186/1471-2105-5-100

Bristow, R. G., Ozcelik, H., Jalali, F., Chan, N., and Vesprini, D. (2007). Homologous recombination and prostate cancer: a model for novel DNA repair targets and therapies. *Radiother. Oncol.* 83 (3), 220–230. doi: 10.1016/j.radonc.2007.04.016

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45 (D1), D369–D379. doi: 10.1093/nar/gkw1102

Chiffoleau, E. (2018). C-type lectin-like receptors as emerging orchestrators of sterile inflammation represent potential therapeutic targets. *Front. Immunol.* 9, 227. doi: 10.3389/fimmu.2018.00227

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140. doi: 10.1038/msb4100180

Collins, K., Jacks, T., and Pavletich, N. P. (1997). The cell cycle and cancer. *Proc. Natl Acad. Sci. U. S. A.* 94 (7), 2776–2778. doi: 10.1073/pnas.94.7.2776

Cox, M. E., Gleave, M. E., Zakikhani, M., Bell, R. H., Piura, E., Vickers, E., et al. (2009). Insulin receptor expression by human prostate cancers. *Prostate* 69 (1), 33–40. doi: 10.1002/pros.20852

Cronin, S. J., and Penninger, J. M. (2007). From T-cell activation signals to signaling control of anti-cancer immunity. *Immunol. Rev.* 220, 151–168. doi: 10.1111/j.1600-065X.2007.00570.x

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter J. Complex Syst.* 1695 (5), 1–9.

Cui, C., Merritt, R., Fu, L., and Pan, Z. (2017). Targeting calcium signaling in cancer therapy. *Acta Pharm. Sin. B* 7 (1), 3–17. doi: 10.1016/j.apsb.2016.11.001

Danielsen, S. A., Eide, P. W., Nesbakken, A., Guren, T., Leithe, E., and Lothe, R. A. (2015). Portrait of the PI3K/AKT pathway in colorectal cancer. *Biochim. Biophys. Acta* 1855 (1), 104–121. doi: 10.1016/j.bbcan.2014.09.008

Del Prete, A., Salvi, V., and Sozzani, S. (2014). Adipokines as potential biomarkers in rheumatoid arthritis. *Mediators Inflamm.* 2014, 425068. doi: 10.1155/2014/425068

Ding, D., Yao, Y., Zhang, S., Su, C., and Zhang, Y. (2017). C-type lectins facilitate tumor metastasis. *Oncol. Lett.* 13 (1), 13–21. doi: 10.3892/ol.2016.5431

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24 (13), i223–i231. doi: 10.1093/bioinformatics/btn161

Doungpan, N., Engchuan, W., Chan, J. H., and Meechai, A. (2016). GSNFS: gene subnetwork biomarker identification of lung cancer expression data. *BMC Med. Genomics* 9 (Suppl 3), 70. doi: 10.1186/s12920-016-0231-4

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4 (8), 1184–1191. doi: 10.1038/nprot.2009.97

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1), 207–210. doi: 10.1093/nar/30.1.207

Edlind, M. P., and Hsieh, A. C. (2014). PI3K-AKT-mTOR signaling in prostate cancer progression and androgen deprivation therapy resistance. *Asian J. Androl.* 16 (3), 378–386. doi: 10.4103/1008-682X.122876

El Sheikh, S. S., Domin, J., Abel, P., Stamp, G., and Lalani el, N. (2003). Androgen-independent prostate cancer: potential role of androgen and ErbB receptor signal transduction crosstalk. *Neoplasia* 5 (2), 99–109. doi: 10.1016/S1476-5586(03)80001-5

Elfiky, A. A., and Jiang, Z. (2013). The PI3 kinase signaling pathway in prostate cancer. *Curr. Cancer Drug Targets* 13 (2), 157–164. doi: 10.2174/1568009611313020005

Emmert-Streib, F., and Glazko, G. V. (2011). Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput. Biol.* 7 (5), e1002053. doi: 10.1371/journal.pcbi.1002053

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The Reactome pathway Knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–d655. doi: 10.1093/nar/gkx1132

Fajardo, A. M., Piazza, G. A., and Tinsley, H. N. (2014). The role of cyclic nucleotide signaling pathways in cancer: targets for prevention and treatment. *Cancers (Basel)* 6 (1), 436–458. doi: 10.3390/cancers6010436

Fang, J. Y., and Richardson, B. C. (2005). The MAPK signalling pathways and colorectal cancer. *Lancet Oncol.* 6 (5), 322–327. doi: 10.1016/S1470-2045(05)70168-6

Fang, S., and Fang, X. (2016). Advances in glucose metabolism research in colorectal cancer. *Biomed. Rep.* 5 (3), 289–295. doi: 10.3892/br.2016.719

Farhan, M., Wang, H., Gaur, U., Little, P. J., Xu, J., and Zheng, W. (2017). FOXO signaling pathways as therapeutic targets in cancer. *Int. J. Biol Sci.* 13 (7), 815–827. doi: 10.7150/ijbs.20052

Farrow, J. M., Yang, J. C., and Evans, C. P. (2014). Autophagy as a modulator and target in prostate cancer. *Nat. Rev. Urol.* 11 (9), 508–516. doi: 10.1038/nrurol.2014.196

Fortney, K., Kotlyar, M., and Jurisica, I. (2010). Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biol.* 11 (2), R13. doi: 10.1186/gb-2010-11-2-r13

Francipane, M. G., and Lagasse, E. (2014). mTOR pathway in colorectal cancer: an update. *Oncotarget* 5 (1), 49–66. doi: 10.18632/oncotarget.1548

Frommer, K. W., Neumann, E., and Müller-Ladner, U. (2011). Adipocytokines and autoimmunity. *Arthritis Res. Ther.* 13 (Suppl 2), O8–O8. doi: 10.1186/ar3412

García-Barros, M., Coant, N., Truman, J.-P., Snider, A. J., and Hannun, Y. A. (2014). Sphingolipids in colon cancer. *Biochim. Biophys. Acta* 1841 (5), 773–782. doi: 10.1016/j.bbalip.2013.09.007

Gkika, D., and Prevarskaya, N. (2011). TRP channels in prostate cancer: the good, the bad and the ugly? *Asian J. Androl.* 13 (5), 673–676. doi: 10.1038/aja.2011.18

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)* 28 (18), i451–i457. doi: 10.1093/bioinformatics/bts389

Goffin, V., Hoang, D. T., Bogorad, R. L., and Nevalainen, M. T. (2011). Prolactin regulation of the prostate gland: a female player in a male game. *Nat. Rev. Urol.* 8 (11), 597–607. doi: 10.1038/nrurol.2011.143

Gomez-Cambronero, J. (2014). Phosphatidic acid, phospholipase D and tumorigenesis. *Adv. Biol. Regul.* 54, 197–206. doi: 10.1016/j.jbior.2013.08.006

Goncalves, P., and Martel, F. (2013). Butyrate and colorectal cancer: the role of butyrate transport. *Curr. Drug. Metab.* 14 (9), 994–1008. doi: 10.2174/1389200211314090006

Gonnissen, A., Isebaert, S., and Haustermans, K. (2013). Hedgehog signaling in prostate cancer and its therapeutic implication. *Int. J. Mol. Sci.* 14 (7), 13979–14007. doi: 10.3390/ijms140713979

Gründker, C., and Emons, G. (2017). The role of gonadotropin-releasing hormone in cancer cell proliferation and metastasis. *Front. Endocrinol.* 8, 187. doi: 10.3389/fendo.2017.00187

Guo, Z., Wang, L., Li, Y., Gong, X., Yao, C., Ma, W., et al. (2007). Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* 23 (16), 2121–2128. doi: 10.1093/bioinformatics/btm294

Gwinner, F., Boulday, G., Vandiedonck, C., Arnould, M., Cardoso, C., Nikolayeva, I., et al. (2017). Network-based analysis of omics data: the LEAN method. *Bioinformatics* 33 (5), 701–709. doi: 10.1093/bioinformatics/btw676

Hagland, H. R., Berg, M., Jolma, I. W., Carlsen, A., and Soreide, K. (2013). Molecular pathways and cellular metabolism in colorectal cancer. *Dig. Surg.* 30 (1), 12–25. doi: 10.1159/000347166

Hartwell, L. H., and Kastan, M. B. (1994). Cell cycle control and cancer. *Science* 266 (5192), 1821–1828. doi: 10.1126/science.7997877

Hassfeld, W., Steiner, G., Studnicka-Benke, A., Skriner, K., Graninger, W., Fischer, I., et al. (1995). Autoimmune response to the spliceosome. an immunologic link between rheumatoid arthritis, mixed connective tissue disease, and systemic lupus erythematosus. *Arthritis Rheum.* 38 (6), 777–785. doi: 10.1002/art.1780380610

He, C. F. P., Su, H., Gu, A., Yan, Z., and Zhu, X. (2017). Disrupted Th1/Th2 balance in patients with rheumatoid arthritis (RA). *Int. J. Clin. Exp. Pathol.* 10 (2), 1233–1242.

Hitchon, C. A., and El-Gabalawy, H. S. (2004). Oxidation in rheumatoid arthritis. *Arthritis Res. Ther.* 6 (6), 265–278. doi: 10.1186/ar1447

Hollande, F., and Papin, M., (2013). "Tight junctions in colorectal cancer" in *Tight junctions in cancer metastasis*. Eds. T. A. Martin and W. G. Jiang (Dordrecht: Springer Netherlands), 149–167. doi: 10.1007/978-94-007-6028-8_7

Housa, D., Housova, J., Vernerova, Z., and Haluzik, M. (2006). Adipocytokines and cancer. *Physiol. Res.* 55 (3), 233–244.

Hsing, A. W., Gao, Y. T., Chua, S., Jr., Deng, J., and Stanczyk, F. Z. (2003). Insulin resistance and prostate cancer risk. *J. Natl. Cancer Inst.* 95 (1), 67–71. doi: 10.1093/jnci/95.1.67

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4 (1), 44–57. doi: 10.1038/nprot.2008.211

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 8 (9), R183. doi: 10.1186/gb-2007-8-9-r183

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1, S233–S240. doi: 10.1093/bioinformatics/18.suppl_1.S233

Iozzo, R. V., and Sanderson, R. D. (2011). Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. *J. Cell Mol. Med.* 15 (5), 1013–1031. doi: 10.1111/j.1582-4934.2010.01236.x

Jarry, A., Charrier, L., Bou-Hanna, C., Devilder, M. C., Crussaire, V., Denis, M. G., et al. (2004). Position in cell cycle controls the sensitivity of colon cancer cells to nitric oxide-dependent programmed cell death. *Cancer Res.* 64 (12), 4227–4234. doi: 10.1158/0008-5472.CAN-04-0254

Jin, R. J., Lho, Y., Connelly, L., Wang, Y., Yu, X., Saint Jean, L., et al. (2008). The nuclear factor-kappaB pathway controls the progression of prostate cancer to androgen-independent growth. *Cancer Res.* 68 (16), 6762–6769. doi: 10.1158/0008-5472.CAN-08-0107

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi: 10.1093/nar/28.1.27

Karni, S., Soreq, H., and Sharan, R. (2009). A network-based method for predicting disease-causing genes. *J. Comput. Biol.* 16 (2), 181–189. doi: 10.1089/cmb.2008.05TT

Kearney, C. J., Cullen, S. P., Tynan, G. A., Henry, C. M., Clancy, D., Lavelle, E. C., et al. (2015). Necroptosis suppresses inflammation via termination of TNF- or LPS-induced cytokine and chemokine production. *Cell Death Differ.* 22 (8), 1313–1327. doi: 10.1038/cdd.2014.222

Khan, A. S., and Frigo, D. E. (2017). A spatiotemporal hypothesis for the regulation, role, and targeting of AMPK in prostate cancer. *Nat. Rev. Urol.* 14 (3), 164–180. doi: 10.1038/nrurol.2016.272

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8 (2), e1002375. doi: 10.1371/journal.pcbi.1002375

Klammer, M., Godl, K., Tebbe, A., and Schaab, C. (2010). Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics* 11, 351. doi: 10.1186/1471-2105-11-351

Knights, A. J., Funnell, A. P. W., Crossley, M., and Pearson, R. C. M. (2012). Holding tight: cell junctions and cancer spread. *Trends Cancer Res.* 8, 61–69.

Knutson, K. L., and Disis, M. L. (2005). Tumor antigen-specific T helper cells in cancer immunity and immunotherapy. *Cancer Immunol. Immunother.* 54 (8), 721–728. doi: 10.1007/s00262-004-0653-2

Lee, S. H., Chang, D. K., Goel, A., Boland, C. R., Bugbee, W., Boyle, D. L., et al. (2003). Microsatellite instability and suppressed DNA repair enzyme expression in rheumatoid arthritis. *J. Immunol.* 170 (4), 2214–2220. doi: 10.4049/jimmunol.170.4.2214

Leipe, J., Grunke, M., Dechant, C., Reindl, C., Kerzendorf, U., Schulze-Koops, H., et al. (2010). Role of Th17 cells in human autoimmune arthritis. *Arthritis Rheum.* 62 (10), 2876–2885. doi: 10.1002/art.27622

Li, S., Yu, Y., Yue, Y., Zhang, Z., and Su, K. (2013). Microbial infection and rheumatoid arthritis. *J. Clin. Cell Immunol.* 4 (6), 174. doi: 10.4172/2155-9899.1000174

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27 (12), 1739–1740. doi: 10.1093/bioinformatics/btr260

Liu, H., and Pope, R. M. (2003). The role of apoptosis in rheumatoid arthritis. *Curr. Opin. Pharmacol.* 3 (3), 317–322. doi: 10.1016/S1471-4892(03)00037-7

Liu, L., Wei, J., and Ruan, J. (2017a). Pathway enrichment analysis with networks. *Genes* 8 (10), 246. doi: 10.3390/genes8100246

Liu, M., Liberzon, A., Kong, S. W., Lai, W. R., Park, P. J., Kohane, I. S., et al. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* 3 (6), e96–e96. doi: 10.1371/journal.pgen.0030096

Liu, T., Zhang, L., Joo, D., and Sun, S.-C. (2017b). NF-κB signaling in inflammation. *Signal Transduction Targeted Ther.* 2, 17023. doi: 10.1038/sigtrans.2017.23

Löffler, I., Grün, M., Böhmer, F. D., and Rubio, I. (2008). Role of cAMP in the promotion of colorectal cancer cell growth by prostaglandin E2. *BMC Cancer* 8, 380–380. doi: 10.1186/1471-2407-8-380

Loo, Y.-M., and Gale, M., Jr. (2011). Immune signaling by RIG-I-like receptors. *Immunity* 34 (5), 680–692. doi: 10.1016/j.immuni.2011.05.003

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29 (14), 1830–1831. doi: 10.1093/bioinformatics/btt285

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 27 (9), 1290–1298. doi: 10.1093/bioinformatics/btr136

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies

(GWAS Catalog). *Nucleic Acids Res.* 45 (D1), D896–D901. doi: 10.1093/nar/gkw1133

Makarov, S. S. (2001). NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Res.* 3 (4), 200–206. doi: 10.1186/ar300

Malemud, C. J. (2013). Intracellular signaling pathways in rheumatoid arthritis. *J. Clin. Cell Immunol.* 4, 160. doi: 10.4172/2155-9899.1000160

Malemud, C. J. (2015). The PI3K/Akt/PTEN/mTOR pathway: a fruitful target for inducing cell death in rheumatoid arthritis? *Future Med. Chem.* 7 (9), 1137–1147. doi: 10.4155/fmc.15.55

Malemud, C. J. (2018). The role of the JAK/STAT signal pathway in rheumatoid arthritis. Therapeutic advances in musculoskeletal disease. *Ther. Adv. Musculoskelet. Dis.* 10, 5–6, 117–127. doi: 10.1177/1759720X18776224

Matsuda, S., Hammaker, D., Topolewski, K., Briegel, K. J., Boyle, D. L., Dowdy, S., et al. (2017). Regulation of the cell cycle and inflammatory arthritis by the transcription cofactor LBH gene. *J. Immunol.* 199 (7), 2316–2322. doi: 10.4049/jimmunol.1700719

Maziveyi, M., and Alahari, S. K. (2017). Cell matrix adhesions in cancer: the proteins that form the glue. *Oncotarget* 8 (29), 48471–48487. doi: 10.18632/oncotarget.17265

McCormack, W. J., Parker, A. E., and O'Neill, L. A. (2009). Toll-like receptors and NOD-like receptors in rheumatic diseases. *Arthritis Res. Ther.* 11 (5), 243–243. doi: 10.1186/ar2729

Moradi-Marjaneh, R., Hassanian, S. M., Fiuji, H., Soleimanpour, S., Ferns, G. A., Avan, A., et al. (2018). Toll like receptor signaling pathway as a potential therapeutic target in colorectal cancer. *J. Cell Physiol.* 233 (8), 5613–5622. doi: 10.1002/jcp.26273

Mosesson, Y., Mills, G. B., and Yarden, Y. (2008). Derailed endocytosis: an emerging feature of cancer. *Nat. Rev. Cancer* 8 (11), 835–850. doi: 10.1038/nrc2521

Murillo-Garzon, V., and Kypta, R. (2017). WNT signalling in prostate cancer. *Nat. Rev. Urol.* 14 (11), 683–696. doi: 10.1038/nrurol.2017.144

Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics* 23 (7), 850–858. doi: 10.1093/bioinformatics/btm019

Nikolayeva, I., Guitart Pla, O., and Schwikowski, B. (2018). Network module identification—a widespread theoretical bias and best practices. *Methods* 132, 19–25. doi: 10.1016/j.ymeth.2017.08.008

Nishimura, D. (2001). BioCarta. *Biotech. Software Internet Rep.* 2 (3), 117–120. doi: 10.1089/152791601750294344

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42 (Database issue), D358–D363. doi: 10.1093/nar/gkt1115

Ozisik, O., Bakir-Gungor, B., Diri, B., and Sezerman, O. U. (2017). Active Subnetwork GA: a two stage genetic algorithm approach to active subnetwork search. *Curr. Bioinformatics* 12 (4), 320–328. doi: 10.2174/1574893611666160527100444

Pernis, A. B. (2009). Th17 cells in rheumatoid arthritis and systemic lupus erythematosus. *J. Intern. Med.* 265 (6), 644–652. doi: 10.1111/j.1365-2796.2009.02099.x

Pickup, M. W., Mouw, J. K., and Weaver, V. M. (2014). The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* 15 (12), 1243–1253. doi: 10.15252/embr.201439246

Pitzalis, C., Kingsley, G., and Panayi, G. (1994). Adhesion molecules in rheumatoid arthritis: role in the pathogenesis and prospects for therapy. *Ann. Rheum. Dis.* 53 (5), 287–288. doi: 10.1136/ard.53.5.287

Powell, J. A. (2014). GO2MSIG, an automated GO based multi-species gene set generator for gene set enrichment analysis. *BMC Bioinformatics* 15, 146. doi: 10.1186/1471-2105-15-146

Qiu, Y. Q., Zhang, S., Zhang, X. S., and Chen, L. (2009). Identifying differentially expressed pathways via a mixed integer linear programming model. *IET Syst. Biol.* 3 (6), 475–486. doi: 10.1049/iet-syb.2008.0155

Quiñonez-Flores, C. M., González-Chávez, S. A., and Pacheco-Tena, C. (2016). Hypoxia and its implications in rheumatoid arthritis. *J. Biomed. Sci.* 23 (1), 62–62. doi: 10.1186/s12929-016-0281-0

Remans, P., Reedquist, K., Bos, J., Verweij, C., Breedveld, F., van Laar, J., et al. (2002). Deregulated Ras and Rap1 signaling in rheumatoid arthritis T cells leads to persistent production of free radicals. *Arthritis Res.* 4 (Suppl 1), 52. doi: 10.1186/ar495

Rihl, M., Kruithof, E., Barthel, C., De Keyser, F., Veys, E. M., Zeidler, H., et al. (2005). Involvement of neurotrophins and their receptors in spondyloarthritis synovitis: relation to inflammation and response to treatment. *Ann. Rheum. Dis.* 64 (11), 1542–1549. doi: 10.1136/ard.2004.032599

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi: 10.1093/nar/gkv007

Rodríguez-Berriguete, G., Fraile, B., Martínez-Onsurbe, P., Olmedilla, G., Paniagua, R., and Royuela, M. (2012). MAP kinases and prostate cancer. *J. Signal Transduct.* 2012, 1–9. doi: 10.1155/2012/169170

Sackmann-Sala, L., and Goffin, V. (2015). Prolactin-induced prostate tumorigenesis. *Adv. Exp. Med. Biol.* 846, 221–242. doi: 10.1007/978-3-319-12114-7_10

Santhanam, S., Alvarado, D. M., and Ciorba, M. A. (2016). Therapeutic targeting of inflammation and tryptophan metabolism in colon and gastrointestinal cancer. *Transl. Res.* 167 (1), 67–79. doi: 10.1016/j.trsl.2015.07.003

Saxena, M., and Yeretssian, G. (2014). NOD-like receptors: master regulators of inflammation and cancer. *Front. Immunol.* 5, 327–327. doi: 10.3389/fimmu.2014.00327

Segui, B., Andrieu-Abadie, N., Jaffrezou, J. P., Benoist, H., and Levade, T. (2006). Sphingolipids as modulators of cancer cell death: potential therapeutic targets. *Biochim. Biophys. Acta* 1758 (12), 2104–2120. doi: 10.1016/j.bbamem.2006.05.024

Shaw, J., Costa-Pinheiro, P., Patterson, L., Drews, K., Spiegel, S., and Kester, M. (2018). Novel sphingolipid-based cancer therapeutics in the personalized medicine era. *Advances in Cancer Research Sphingolipids in Cancer* 327–366. doi: 10.1016/bs.acr.2018.04.016

Shrivastav, M., Mittal, B., Aggarwal, A., and Misra, R. (2002). Autoantibodies against cytoskeletal proteins in rheumatoid arthritis. *Clin. Rheumatol.* 21 (6), 505–510. doi: 10.1007/s100670200124

Shukla, S., Maclennan, G. T., Hartman, D. J., Fu, P., Resnick, M. I., and Gupta, S. (2007). Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion. *Int. J. Cancer* 121 (7), 1424–1432. doi: 10.1002/ijc.22862

Siddiqui, N., and Borden, K. L. (2012). mRNA export and cancer. *Wiley Interdiscip. Rev. RNA* 3 (1), 13–25. doi: 10.1002/wrna.101

Silvertown, J. D., Summerlee, A. J., and Klonisch, T. (2003). Relaxin-like peptides in cancer. *Int. J. Cancer* 107 (4), 513–519. doi: 10.1002/ijc.11424

Slattery, M. L., Lundgreen, A., Kadlubar, S. A., Bondurant, K. L., and Wolff, R. K. (2013). JAK/STAT/SOCS-signaling pathway and colon and rectal cancer. *Mol. Carcinog.* 52 (2), 155–166. doi: 10.1002/mc.21841

Slattery, M. L., Mullany, L. E., Wolff, R. K., Sakoda, L. C., Samowitz, W. S., and Herrick, J. S. (2018). The p53-signaling pathway and colorectal cancer: interactions between downstream p53 target genes and miRNAs. *Genomics* 111 (4), 762–771. doi: 10.1016/j.ygeno.2018.05.006

Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics* 20 (10), 1517–1521. doi: 10.1093/bioinformatics/bth112

Stacker, S. A., and Achen, M. G. (2013). The VEGF signaling pathway in cancer: the road ahead. *Chin. J. Cancer* 32 (6), 297–302. doi: 10.5732/cjc.012.10319

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34 (Database issue), D535–D539. doi: 10.1093/nar/gkj109

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102

Sumitomo, S., Nagafuchi, Y., Tsuchida, Y., Tsuchiya, H., Ota, M., Ishigaki, K., et al. (2018). A gene module associated with dysregulated TCR signaling pathways in CD4(+) T cell subsets in rheumatoid arthritis. *J. Autoimmun.* 89, 21–29. doi: 10.1016/j.jaut.2017.11.001

Sun, W. (2012). Angiogenesis in metastatic colorectal cancer and the benefits of targeted therapy. *J. Hematol. Oncol.* 5, 63–63. doi: 10.1186/1756-8722-5-63

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J., et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics* 25 (1), 75–82. doi: 10.1093/bioinformatics/btn577

Ulgen, E., Ozisik, O., and Sezerman, O. U. (2018). pathfindR: an R package for pathway enrichment analysis utilizing active subnetworks. *bioRxiv* 272450. doi: 10.1101/272450

Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* 1, 8. doi: 10.1186/1752-0509-1-8

Ulitsky, I., and Shamir, R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics* 25 (9), 1158–1164. doi: 10.1093/bioinformatics/btp118

Vasilopoulos, Y., Gkretsi, V., Armaka, M., Aidinis, V., and Kollias, G. (2007). Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in TNF-driven arthritis. *Ann. Rheum. Dis.* 66 (Suppl 3), iii23–iii28. doi: 10.1136/ard.2007.079822

Wang, J., and Maldonado, M. A. (2006). The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cell Mol. Immunol.* 3 (4), 255–261.

Wang, J., Xu, K., Wu, J., Luo, C., Li, Y., Wu, X., et al. (2012). The changes of Th17 cells and the related cytokines in the progression of human colorectal cancers. *BMC Cancer* 12, 418. doi: 10.1186/1471-2407-12-418

Wang, X. X., Xiao, F. H., Li, Q. G., Liu, J., He, Y. H., and Kong, Q. P. (2017). Large-scale DNA methylation expression analysis across 12 solid cancers reveals hypermethylation in the calcium-signaling pathway. *Oncotarget* 8 (7), 11868–11876. doi: 10.18632/oncotarget.14417

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38 (Web Server issue), W214–W220. doi: 10.1093/nar/gkq537

Watson, A. J. M. (2004). Apoptosis and colorectal cancer. *Gut* 53 (11), 1701–1709. doi: 10.1136/gut.2004.052704

Weber, M. J., and Gioeli, D. (2004). Ras signaling in prostate cancer progression. *J. Cell Biochem.* 91 (1), 13–25. doi: 10.1002/jcb.10683

Wen, Y. A., Xing, X., Harris, J. W., Zaytseva, Y. Y., Mitov, M. I., Napier, D. L., et al. (2017). Adipocytes activate mitochondrial fatty acid oxidation and autophagy to promote tumor growth in colon cancer. *Cell Death Dis.* 8 (2), e2593. doi: 10.1038/cddis.2017.21

Werner, T. (2008). Bioinformatics applications for pathway analysis of microarray data. *Curr. Opin. Biotechnol.* 19 (1), 50–54. doi: 10.1016/j.copbio.2007.11.005

Whitaker, H. C., and Neal, D. E. (2010). RAS pathways in prostate cancer—mediators of hormone resistance? *Curr. Cancer Drug Targets* 10 (8), 834–839. doi: 10.2174/156800910793358005

Wierzbicki, P. M., and Rybarczyk, A. (2015). The Hippo pathway in colorectal cancer. *Folia Histochem. Cytobiol.* 53 (2), 105–119. doi: 10.5603/FHC.a2015.0015

Wu, D., Wu, P., Huang, Q., Liu, Y., Ye, J., and Huang, J. (2013). Interleukin-17: a promoter in colorectal cancer progression. *Clin. Dev. Immunol.* 2013, 436307–436307. doi: 10.1155/2013/436307

Wu, J., Gan, M., and Jiang, R. (2011). A genetic algorithm for optimizing subnetwork markers for the study of breast cancer metastasis. 2011 Seventh International Conference on Natural Computation, USA: IEEE. 3, 1578–1582. doi: 10.1109/ICNC.2011.6022270

Wu, X., Hasan, M. A., and Chen, J. Y. (2014). Pathway and network analysis in proteomics. *J. Theor. Biol.* 362, 44–52. doi: 10.1016/j.jtbi.2014.05.031

Wysocka, M. B., Pietraszek-Gremplewicz, K., and Nowak, D. (2018). The role of apelin in cardiovascular diseases, obesity and cancer. *Front. Physiol.* 9, 557–557. doi: 10.3389/fphys.2018.00557

Yamaguchi, H., and Condeelis, J. (2007). Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochim. Biophys. Acta* 1773 (5), 642–652. doi: 10.1016/j.bbamcr.2006.07.001

Yan, H., Kamiya, T., Suabjakyong, P., and Tsuji, N. M. (2015). Targeting C-type lectin receptors for cancer immunity. *Front. Immunol.* 6, 408. doi: 10.3389/fimmu.2015.00408

Yang, X. Y., Zheng, K. D., Lin, K., Zheng, G., Zou, H., Wang, J. M., et al. (2015). Energy metabolism disorder as a contributing factor of rheumatoid arthritis: a comparative proteomic and metabolomic study. *PLoS One* 10 (7), e0132695. doi: 10.1371/journal.pone.0132695

Yin, B., Liu, W., Yu, P., Liu, C., Chen, Y., Duan, X., et al. (2017). Association between human papillomavirus and prostate cancer: a meta-analysis. *Oncol. Lett.* 14 (2), 1855–1865. doi: 10.3892/ol.2017.6367

You, M., Yuan, S., Shi, J., and Hou, Y. (2015). PPARdelta signaling regulates colorectal cancer. *Curr. Pharm. Des.* 21 (21), 2956–2959. doi: 10.2174/1381612821666150514104035

Zenonos, K., and Kyprianou, K. (2013). RAS signaling pathways, mutations and their role in colorectal cancer. *World J. Gastrointest. Oncol.* 5 (5), 97–101. doi: 10.4251/wjgo.v5.i5.97

Zhan, T., Rindtorff, N., and Boutros, M. (2017). Wnt signaling in cancer. *Oncogene* 36 (11), 1461–1473. doi: 10.1038/onc.2016.304

Zhang, L., Yang, S., Chen, X., Stauffer, S., Yu, F., Lele, S. M., et al. (2015). The hippo pathway effector YAP regulates motility, invasion, and castration-resistant growth of prostate cancer cells. *Mol. Cell Biol.* 35 (8), 1350–1362. doi: 10.1128/MCB.00102-15

Zhang, T., Ma, Y., Fang, J., Liu, C., and Chen, L. (2017a). A deregulated PI3K-AKT signaling pathway in patients with colorectal cancer. *J. Gastrointest. Cancer* 50 (1), 35–41. doi: 10.1007/s12029-017-0024-9

Zhang, Y. L., Wang, R. C., Cheng, K., Ring, B. Z., and Su, L. (2017b). Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer Biol. Med.* 14 (1), 90–99. doi: 10.20892/j.issn.2095-3941.2016.0086

Zhao, P., and Zhang, Z. (2018). TNF-α promotes colon cancer cell migration and invasion by upregulating TROP-2. *Oncol. Lett.* 15 (3), 3820–3827. doi: 10.3892/ol.2018.7735

Zhao, S., Zhang, Y., Zhang, Q., Wang, F., and Zhang, D. (2014). Toll-like receptors and prostate cancer. *Front. Immunol.* 5, 352. doi: 10.3389/fimmu.2014.00352

Zhao, X. M., Wang, R. S., Chen, L., and Aihara, K. (2008). Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res.* 36 (9), e48. doi: 10.1093/nar/gkn145

Zhong, M., Boseman, M. L., Millena, A. C., and Khan, S. A. (2010). Oxytocin induces the migration of prostate cancer cells: involvement of the Gi-coupled signaling pathway. *Mol. Cancer Res.* 8 (8), 1164–1172. doi: 10.1158/1541-7786.MCR-09-0329

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.

Copyright © 2019 Ulgen, Ozisik and Sezerman. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 13 March 2019
doi: 10.3389/fgene.2019.00164

Adapting Community Detection Algorithms for Disease Module Identification in Heterogeneous Biological Networks

Beethika Tripathi[1,2,3], *Srinivasan Parthasarathy*[4,5], *Himanshu Sinha*[2,3,6], *Karthik Raman*[2,3,6] *and Balaraman Ravindran*[1,2,3]*

[1] Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, [2] Initiative for Biological Systems Engineering, Indian Institute of Technology Madras, Chennai, India, [3] Robert Bosch Centre for Data Science and AI, Indian Institute of Technology Madras, Chennai, India, [4] Department of Computer Science and Engineering, Ohio State University, Columbus, OH, United States, [5] Department of Biomedical Informatics, Ohio State University, Columbus, OH, United States, [6] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India

OPEN ACCESS

Edited by:
Marco Antoniotti,
University of Milano-Bicocca, Italy

Reviewed by:
Zhi-Ping Liu,
Shandong University, China
Xianwen Ren,
Peking University, China

***Correspondence:**
Balaraman Ravindran
ravi@cse.iitm.ac.in

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 01 November 2018
Accepted: 14 February 2019
Published: 13 March 2019

Citation:
Tripathi B, Parthasarathy S, Sinha H, Raman K and Ravindran B (2019) Adapting Community Detection Algorithms for Disease Module Identification in Heterogeneous Biological Networks. Front. Genet. 10:164. doi: 10.3389/fgene.2019.00164

Biological networks catalog the complex web of interactions happening between different molecules, typically proteins, within a cell. These networks are known to be highly modular, with groups of proteins associated with specific biological functions. Human diseases often arise from the dysfunction of one or more such proteins of the biological functional group. The ability, to identify and automatically extract these modules has implications for understanding the etiology of different diseases as well as the functional roles of different protein modules in disease. The recent DREAM challenge posed the problem of identifying disease modules from six heterogeneous networks of proteins/genes. There exist many community detection algorithms, but all of them are not adaptable to the biological context, as these networks are densely connected and the size of biologically relevant modules is quite small. The contribution of this study is 3-fold: first, we present a comprehensive assessment of many classic community detection algorithms for biological networks to identify non-overlapping communities, and propose heuristics to identify small and structurally well-defined communities—*core* modules. We evaluated our performance over 180 GWAS datasets. In comparison to traditional approaches, with our proposed approach we could identify 50% more number of disease-relevant modules. Thus, we show that it is important to identify more compact modules for better performance. Next, we sought to understand the peculiar characteristics of disease-enriched modules and what causes standard community detection algorithms to detect so few of them. We performed a comprehensive analysis of the interaction patterns of known disease genes to understand the structure of disease modules and show that merely considering the known disease genes set as a module does not give good quality clusters, as measured by typical metrics such as modularity and conductance. We go on to present a methodology leveraging these known disease genes, to also include the neighboring nodes of these genes into

a module, to form good quality clusters and subsequently extract a "gold-standard set" of disease modules. Lastly, we demonstrate, with justification, that "overlapping" community detection algorithms should be the preferred choice for disease module identification since several genes participate in multiple biological functions.

Keywords: overlapping community detection, non-overlapping community detection, disease module identification, biological networks, heterogeneous networks

1. INTRODUCTION

Biological networks, such as protein–protein interaction networks, gene regulatory networks, gene co-expression networks, metabolic networks, signaling networks provide a mathematical representation of biological systems. In this work, we are interested in the study of networks that encode interactions among proteins. These interactions can be *physical*, where proteins bind to one another, or *functional*, where proteins are associated with one another for performing a particular task. Analyzing biological networks is essential for guiding biological experiments—these experiments could otherwise be very difficult to perform, or even intractable, if every gene or protein were to be characterized individually.

Biological networks have been observed to be highly modular (Hartwell et al., 1999), where a tightly connected group of genes (nodes) are involved in similar biological functions. These groups are referred to as communities, modules, or clusters. Modules detected from biological networks are usually responsible for a common phenotype and are useful in providing insights pertaining to biological functionality. Module identification methods (also known as community detection methods) play a crucial role in obtaining these functional modules.

Disease phenotypes are usually caused by the malfunctioning of certain genes, these group of genes is referred to as *disease module*. As genes responsible for a phenotype often possess common functionality, there exists a strong association between disease modules and functional modules (Goh et al., 2007; Zanzoni et al., 2009; Barabási et al., 2011). We know that the modular structure of the biological network is often useful in identifying functional modules; so, we proceed with the assumption that the same would be useful to identify disease modules. It is essential to identify these disease modules, as it could be helpful for various applications, such as the comprehensive molecular understanding of the disease, identification of co-occurring diseases, or the identification of extensive set of genes for targeted therapies.

Present Work. Various algorithms exist in the literature for community detection (module identification). Many are evaluated on *in silico* generated benchmark networks (Friedman et al., 2001; Girvan and Newman, 2002; Newman, 2006). However, performance of these multitude of community detection approaches across variety of these biological networks to discover biologically relevant modules (disease modules or functional modules) remains poorly understood. Such a diverse set of biological networks are fundamentally different owing to the generative processes underpinning their structure, it is important to evaluate performance of different approaches across them. In this work, we study the adaptability of these community detection approaches for disease module identification, notably in the context of the recent an open-community challenge called as the DREAM challenge (Dialogue for Reverse Engineering Assessments and Methods) on Disease Module Identification (DMI)[1]. The challenge posed the problem of predicting "non-overlapping" and small modules of size ranging from 3 to 100 nodes, across six different networks. The set of predicted modules from a community detection method were evaluated against 180 GWAS datasets to find out any significant association of modules with complex trait or disease, to identify disease modules amongst them.

We comprehensively assessed various existing module identification algorithms across diverse biological networks and propose novel algorithms with the notion of *core communities*, to identify small and structurally well-defined communities. We obtained a substantial improvement over the traditional approaches. To our concern, a common problem existed for all the non-overlapping clustering approaches—the number of enriched modules were quite less in comparison to the number of modules predicted. Also, the number of diseases enriching the modules were very less in comparison to the number of different GWAS datasets (180 GWAS datasets) available for testing. These observations beg multiple questions: (1) Does the disease module possess a community structure at all? (2) Could we build "ground-truth disease modules" whose structural properties could be analyzed? (3) Do all of the diseases have structurally well-defined modules associated with them? (4) Most importantly, is "non-overlapping" community detection suitable for disease module identification as in this challenge? (5) Lastly, is there any association between the diseases, in terms of common nodes in the community structure?

We address all of these questions in the present study. In summary, our main contributions are as follows:

- We have introduced a framework for *core module identification*, to identify small and structurally well-defined communities. We show that this is important to identify compact modules from biological networks, and to achieve a better performance in identifying disease-relevant modules.
- We report a comprehensive assessment of many classic community detection algorithms across 6 different types of biological networks, evaluated over 180 GWAS datasets. With our proposed approach, we achieved 50% performance

[1] https://www.synapse.org/#!Synapse:syn6156761/wiki/400645

improvement in identifying disease-relevant modules over classical approaches.
- We have also analyzed the patterns of connectivity in a disease module to better understand the properties of disease modules. We propose a method to identify *gold standard disease modules* based on the genes already shown to be associated with a particular disease.
- We show that *overlapping community detection* is a better approach for the identification of disease-relevant modules. Overlapping community detection is a preferred solution as a gene could be responsible for multiple diseases, and hence should be part of various disease modules.
- We have utilized *overlaps* of the disease modules, which are genes that are involved in multiple diseases (or disease module), to identify diseases that occur together, i.e., *co-morbid diseases*.

2. MATERIALS AND METHODS

2.1. Data

In this section, we summarize the six different biological networks that were made available as part of the DREAM challenge. We have identified disease modules in each of these networks. We also introduce the Genome-Wide Association Study (GWAS) dataset that is central for evaluating the modules predicted by the community detection algorithms.

2.1.1. DREAM Challenge Biological Networks

The organizers of the DMI DREAM challenge provided a unique collection of biological networks for humans. This collection included multiple physical interaction networks (protein interaction networks, signaling network) and functional gene networks (co-expression, homology, and cancer). The statistics on the number of nodes and edges in these networks are presented in **Table 1**. In this section, we will briefly describe these networks.

Protein-Protein Interaction Network-1: The human protein-protein interaction network-1 (PPI-1) data were obtained from STRING version 10.0 (Szklarczyk et al., 2014) after removing the interactions derived from text mining. In this network, the nodes represent proteins, and the edges represent interactions, with the weights representing confidence scores.

Protein-Protein Interaction Network-2: Similar to PPI-1, this is also a protein interaction network, obtained from InWeb (Li et al., 2017), where the interactions are aggregated from primary databases and literature. Again, the proteins are the nodes in the network, and their reported physical interactions are the edges. The edge weights represent the confidence in each interaction.

Signaling Network: Türei et al. (2016) have provided the signaling network, which represents signaling pathways. In this case, the nodes are the genes, and the directed edges between them represent the gene interactions responsible for a cellular function. The weights represent the confidence scores from the experiments that have reported the interaction. Genes in this network can be mapped to corresponding proteins in the other networks.

Co-expression Network: Co-expression network was obtained from Gene Expression Omnibus (Barrett et al., 2010) and captures the correlation between the expression patterns of genes. These expression patterns of genes are observed across various samples of the experiments performed under different experimental conditions. The network is created with genes as nodes and co-expression as the edge between them.

Cancer Network: The cancer network is derived from Project Achilles (Cowley et al., 2014), which performed experiments to determine tumor-wise essential genes that are critical for the survival of that tumor. Those genes that are essential and are absolutely necessary for a tumor to function are connected through an edge in the cancer network. These correlations between the gene expression patterns with respect to a tumor are studied across various tumor samples. The correlation scores obtained through these experiments are represented as edge weights.

Homology Network: The homology network is constructed by connecting genes which are evolutionarily related. Evolutionarily related genes were identified using the Clustering by Inferred Models of Evolution (CLIME) (Li et al., 2014) algorithm. The algorithm partitions the genes into sets of evolutionarily conserved module. The algorithm also provides the confidence scores based on the evolutionary evidence, which are represented as weights of the edges connecting the evolutionarily connected genes in the homology network.

2.1.2. Pre-processing

Biological networks being noisy, pre-processing these networks plays an important role. We sparsified the networks by removing the edges with low weights. We removed edges having weights lesser than two standard deviations from the mean. This not only reduces computation time for the various approaches but also improves the performance of methods by reducing noise.

2.1.3. Genome-Wide Association Study (GWAS)

A Genome-Wide Association Study is an observational study conducted across different individuals. The objective of the study is to investigate the association between genetic variants across the whole genome and traits. The genetic variants refer to the variations that occur in a nucleotide at any specific position in a genome. We have a comprehensive set of 180 GWAS datasets associated with various complex traits and diseases, which belong to broader categories of 15 diseases, as shown in **Table S1**. Modules predicted by the community detection algorithms are tested against each of these GWAS datasets.

2.2. Methods for Module Identification

This section details the various approaches that have been used in our experiments. Methods discussed under "Module identification using non-overlapping community detection" form the basis for our proposed framework, as we detail in section 2.3. Methods discussed under "Overlapping community detection" are primarily used to analyse the properties of disease modules, as discussed in section 3. The purpose of this section is to give an overview of the methods available for module identification in networks, which are leveraged by us to

TABLE 1 | Network statistics of different biological networks.

Network type	#Nodes	#Edges	Edge Weight	Density	Clustering coefficient
PPI-1	17,397	2,232,405	Confidence	0.01475	0.13759
PPI-2	12,420	397,309	Confidence	0.00515	0.12421
Signaling	5,254	21,826	Confidence	0.00133	0.00227
Co-expression	12,588	1,000,000	Correlation	0.01262	0.05209
Cancer	14,679	1,000,000	Correlation	0.00928	0.14288
Homology	10,405	4,223,606	Confidence	0.07803	0.04153

improve module identification in biological networks to identify disease modules.

2.2.1. Module Identification Using Non-overlapping Community Detection

Non-overlapping community detection methods are frequently adopted in the biomedical research (Choobdar et al., 2018). However, such methods restricts every node in a network to belong to a single community, and due to extensive cross talk among various genes and pathways, this restriction in biological networks is untenable. To understand the performance of different module identification methods with such restrictions, we tried some of the most commonly accepted approaches in the field of biology such as modularity maximization (Newman, 2004; Blondel et al., 2008), Markov chain CLustering (MCL) (Dongen, 2000), and Community detection using External and Internal scores in Large networks (CEIL) (Sankar et al., 2015) across various biological networks. We now discuss various state-of-the-art approaches based on (1) quality measures to define community structure, and (2) random-walk based methods to identify community structure.

2.2.1.1. Community quality measures

A network can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of n nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, is a set of e edges. The network are represented using an adjacency matrix A, which is square matrix of dimension $|\mathcal{V}| \times |\mathcal{V}|$. The element A_{ij} in the matrix is zero when there is no edge between node i and node j, and non-zero representing the weight of the edges connecting the nodes; for unweighted networks the value is one. The degree of a node i in the network denoted as d_i, is the number of edges from a node to the other nodes, i.e., $d_i = \sum_{j \in \mathcal{V}} A_{ij}$. Next, we define some important network parameters that enable measurement of community quality.

Modularity: Modularity is defined for a group of nodes, as the difference between the number of edges between those nodes in the original network and a null model, which is essentially a random rewiring of the original network, retaining degree distribution. The higher the difference, the better is the connectivity between the nodes. For a good community the modularity score should be high. The highest value is one. Modularity for a community c is defined as follows:

$$Modularity(c) = \frac{1}{2e} \sum_{i,j \; i \neq j} \left(A_{i,j} - \frac{d_i d_j}{2e} \right) \delta_{c_{(i)} c_{(j)}} \tag{1}$$

where $\frac{d_i d_j}{2e}$ represents the expected number of edges between nodes i and j, $c_{(i)}$ represents the community to which node i belongs and

$$\delta_{c_{(i)} c_{(j)}} = \begin{cases} 1 & \text{if } c_{(i)} = c_{(j)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Modularity based method for community detection prefers group of nodes with higher concentration of edges than expected as communities.

Conductance: Conductance is a measure to define the quality of the community, based on how well-separated the nodes in the community are to the rest of the network. It measures the cut of the community concerning the internal connectivity of the nodes in the network. A good community is isolated from rest of the networks thus have low conductance. The conductance of the community c is defined as:

$$Conductance(c) = \frac{\sum_{i \in c, j \in \bar{c}} A_{i,j}}{min(InternalEdge(c), InternalEdge(\bar{c}))} \tag{3}$$

where $\bar{c_i}$ comprises of the rest of the network other than the nodes in c_i and,

$$InternalEdge(c) = \sum_{i \in c} \sum_{j \in \mathcal{V}} A_{i,j} \tag{4}$$

CEIL: Community detection using External and Internal scores in Large networks (CEIL) (Sankar et al., 2015) is another way of measuring the quality of the community. CEIL strikes the middle ground between modularity and conductance which takes into account: (1) the internal density of the community, and (2) the separability of the community from the rest of the network, measured by internal and external score respectively.

The density of the community is the ratio of internal community edges and possible edges inside the community. The separability of the community from the rest of the network is measured as the ratio of internal community edges and edges that are incident on that community. CEIL Score for a community c

with n_c nodes is the product of internal and external score which are defined below.

$$InternalScore(c) = \begin{cases} \frac{InternalEdge(c)}{\binom{n_c}{2}} & \text{if } n_c > 1 \\ 0 & \text{if } n_c = 0 \end{cases} \quad (5)$$

$$ExternalScore(c) = \frac{InternalEdge(c)}{InternalEdge(c) + \sum_{i \in c, j \in \bar{c}} A_{i,j}} \quad (6)$$

$$CEIL(c) = InternalScore(c) \times ExternalScore(c) \quad (7)$$

2.2.1.2. Markov Chain Clustering

Markov Chain Clustering (MCL) (Dongen, 2000) is a random walk-based approach. With the help of random walks, the flow of the network is analyzed and communities are located where the flow tends to gather. For MCL, two processes are alternated on the network, (1) *expansion*, which involves taking powers of the transition matrix to determine the flow of the network, and (2) *inflation*, which involves re-scaling and normalizing the columns and then taking the power of the column.

The application on real-world network of these methods could be found in the work of Fortunato (2010).

2.2.2. Module Identification Using Overlapping Community Detection

Overlapping community detection allows a node to be part of multiple communities thus resulting in overlapping communities. As genes are commonly involved in multiple functionalities, we have explored overlapping clustering to identify disease modules. The overlapping clustering approaches that we have explored involve two phases to identify communities: (1) "seed node" selection and (2) seed expansion. Since seed node selection is the most critical step to initialize the communities, we have explored multiple strategies to identify nodes that are likely to be "disease nodes." The phases of community detection are discussed below.

2.2.2.1. Seed node selection

We now describe our approach to identify seed nodes, which forms the basis for our algorithm to predict overlapping communities.

Disease seed nodes: Considering the genome-wide significance threshold of 10^{-4} as defined by Choobdar et al. (2018), the genes having a p-value below this threshold were considered as disease seed genes. We also considered 10^{-6} as a threshold to keep a stricter constraint. We defined *disease seed nodes* as the set of genes that pass the threshold across the 180 GWAS datasets.

Unsupervised seed nodes: In the absence of information about known disease nodes, we find a correlation between disease genes and network centrality measures like degree centrality and clustering coefficient of nodes. We observed that disease genes have a higher degree in comparison to the non-disease genes. Consequently, we used HITS (Schütze et al., 2008) and spread hubs (Whang et al., 2016), which are based on the degree of a node, as a seed selection mechanism, to select some important nodes from the network. We grow the communities using PPR scores as described in Andersen et al. (2006). As there is no information involved about the disease seed nodes, we call this process as *unsupervised seed node*.

2.2.2.2. Seed expansion

The seed expansion is done based on the Personalized PageRank (PPR) scores as described in Andersen et al. (2006). PPR scores are used to rank the nodes in the neighborhood of a seed node. The nodes, in the order of their ranking based on PPR scores, are added to the module one by one till the size of the set reaches a particular value (100 for us) as shown in **Figure 1**. The modularity score of the group is computed after the addition of every node. The group of nodes that has the maximum modularity among the different groups, obtained after each addition, forms a module. This seed node expansion process is done for all the seed nodes.

From next section onwards we will discuss about our proposed work.

2.3. Proposed Framework—Core Module Identification

Biological networks exhibit a power-law (Barabási and Albert, 1999) degree distribution, where a few nodes have very high degrees whereas most of the nodes have small degrees. Performing community detection on these networks results in a few giant communities corresponding to the high degree nodes, along with multiple small communities. These giant communities cover most of the network and are least informative to derive any biological insights. Thus, there exists a need to improve the setup to perform community detection. Works in the past, such as those done by Berger and co-workers Singh et al. (2006), take into account the domain knowledge for generating finer clusters. However, identifying finer clusters without any domain knowledge is an interesting problem to be studied. We have proposed few approaches in this section to obtain finer clusters.

We introduce the term *core of the module* to represent finer modules. A core is structurally the strongest part of the module. We have designed four different frameworks to extract the core module, which are explained below:

2.3.1. Ensemble Approach to Clustering

There exist multiple topological definitions of communities and multiple metrics like modularity, conductance, etc. to identify them. However, which topological definition suits a "biologically meaningful" community, is not well-studied. It would be interesting to incorporate multiple topological aspects to generate biologically meaningful modules.

Asur et al. (2007) and subsequent followup (See surveys Ghosh and Acharya, 2013; Ji et al., 2014), suggest ensemble frameworks to combine different clustering algorithms on biological data. Many approaches are suitable for base clustering approaches, that have a fixed number of predicted clusters. Working with a fixed number of clusters might not be the best way of identifying communities from a network, as we do not know *a priori* the desired number of clusters. We develop a simple yet novel approach to compute a consensus from the approaches that do not require the number of clusters to be fixed.

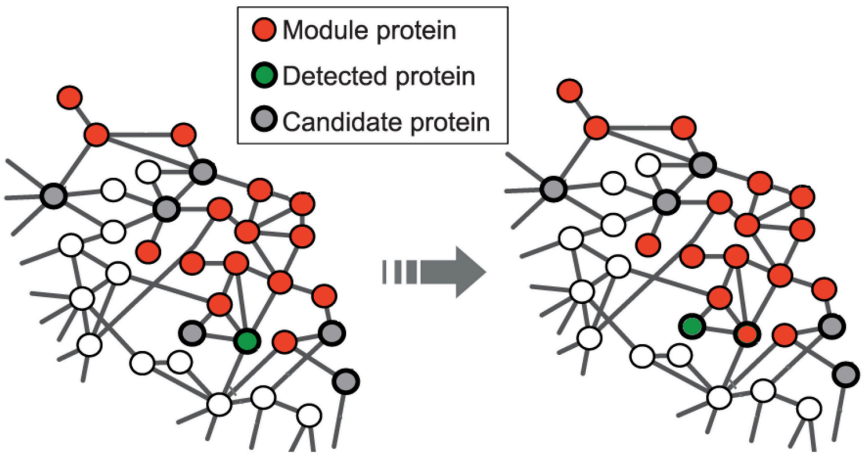


FIGURE 1 | Red nodes represent the module formed by disease seed genes; the set of potential candidate nodes in the local neighborhood of the module is shown in gray; the green node represents the detected gene based on the personalized page rank score and will be included in the module at the next step. This is adapted from the image licensed under the CC BY 4.0 license and attributed to Ghiassian et al. (2015). A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLoS Comput Biol 11(4): e1004120. https://doi.org/10.1371/journal.pcbi.1004120.

We have built an ensemble framework that takes consensus across various approaches, like modularity maximization (with different resistance parameter settings to obtain modules of different sizes), MCL and CEIL. These approaches captures varied aspects of the network structure without fixing the number of clusters to be predicted. We consider r base clustering approaches for a network with a set of nodes $\mathcal{V} = \{v_i\}_{i=1}^{n}$. We build a vector for each node, $\boldsymbol{v}$, $\{[v]_q | \ q = 1, 2, ..r\}$, where each element corresponds to the community assignment of that node in the *qth* clustering algorithm (see **Figure 2A** left).

The pairwise Jaccard similarity (Jaccard, 1901) between nodes, represented as $\boldsymbol{J_{sim}}\{\boldsymbol{v_i}, \boldsymbol{v_j}\} = \frac{\|\{\boldsymbol{v_i} \cap \boldsymbol{v_j}\}\|}{\|\{\boldsymbol{v_i} \cup \boldsymbol{v_j}\}\|}$, is computed to obtain the similarity between the community assignments across all the nodes (as demonstrated in **Figure 2A** right). For example, if the similarity between a given pair of nodes is unity, it means that the nodes co-occurred in the communities predicted by all the algorithms. We then built a similarity matrix out of these pairwise Jaccard similarity values, and subsequently constructed a network from this similarity matrix by linking the nodes having a similarity greater than 0.5. Finally, we use modularity maximization to perform module identification on the resultant network.

2.3.2. Perturbations to Identify Robust Communities

Biological networks have a lot of inherent noise (Bader and Hogue, 2002), caused by the incompleteness of data or experimental biases. Therefore, it is important to identify communities that are robust to noise in the network. To identify robust communities, we follow a setup of perturbing the network multiple times and then performing a community detection on the perturbed networks.

We perturbed the network by randomly dropping 1% of edges. We repeated this for 100 iterations as indicated in **Figures 2B,C**. To detect communities on all the perturbed networks, we follow a setup similar to the ensemble approach described earlier, performing modularity maximization on the similarity network. This enabled us to identify modules *persistent across* perturbed networks. The process is explained in the **Figure 2D**.

2.3.3. Core With Minimum Outgoing Edges

A community should have a higher number of edges connecting the nodes within a community ("internal edges") (Newman, 2004) and a fewer number of edges connecting nodes outside the community ("outgoing edges") (Kannan et al., 2004). For large communities, we identify a core that consists of the nodes that satisfy the property of a good community. These are the nodes that have a higher number of internal edges and a fewer number of outgoing edges. To this end, we have computed a *core score* for each node n, which considers the ratio of outgoing edges to internal edges from that node as follows:

$$CoreScore(n) = \frac{OutgoingEdges(n)}{InternalEdges(n)} \tag{8}$$

We rank the nodes in a module on the basis of their core score, i.e., nodes with lower scores get better ranks. In the case of larger communities of size more than 100, we consider the top 60 nodes as the core and ignore the remaining nodes (**Figure 2F**). We consider only top 60 nodes as we figured out through empirical studies by running multiple experiments that the average size of a disease module is 60. **Figure 3** shows the size distribution of the disease modules obtained using multiple approaches. This approach helped in pruning the least important nodes from modules.

2.3.4. Multiple Core Identification

We defined an iterative way of performing community refinement. In the first step, we used modularity maximization to identify modules in the network. Some of the resulting modules can be quite large due to the high connectivity of

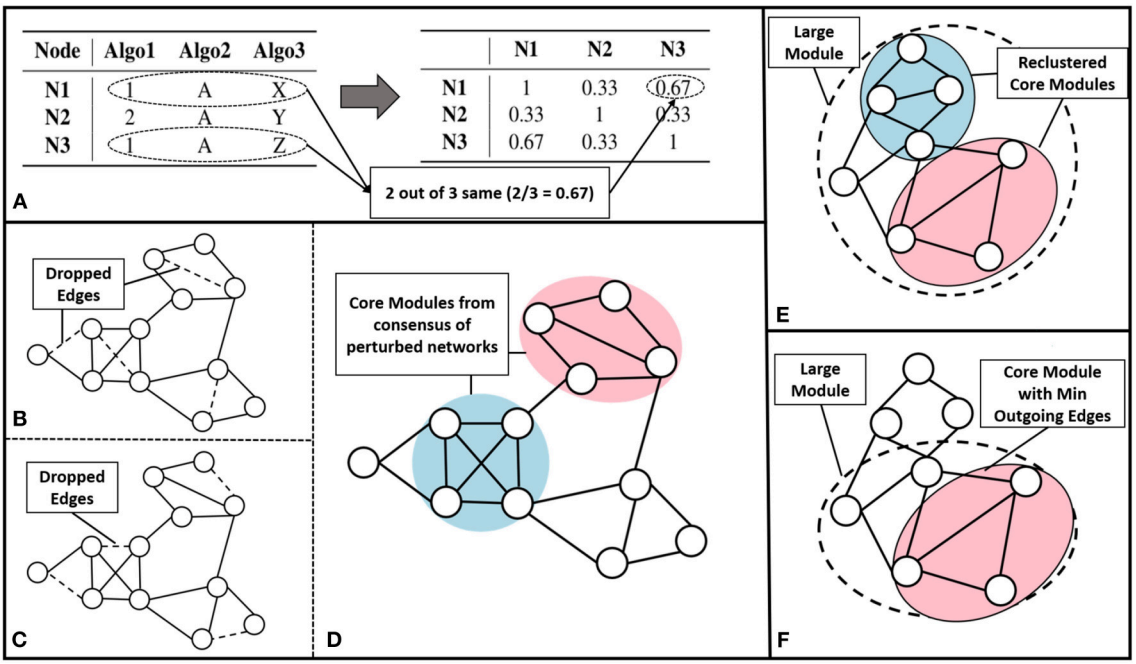


FIGURE 2 | Core module identification methods **(A)** The process of *ensemble clustering*, which uses the power of multiple community detection approaches. A vector of community assignment is created for each node (left). The consensus is taken by computing the Jaccard similarity between the community assignment vectors, for every pair of nodes (right). **(B–D)** *Perturbation*, the process of perturbing the network and finding consensus module across the set of perturbed networks: **(B,C)** are examples of perturbed network after randomly dropping 1% of the edges (dashed lines), **(D)** community detected across the set of 100 perturbed networks. **(E)** *Multiple Core Identification* breaks the large module identified by a community detection algorithm into smaller modules as shown in the example where the dotted circle represents a large module and the colored circles represent the multiple cores obtained, by breaking down the larger modules. **(F)** From a large module *min outgoing edges* selects the group of nodes with minimum outgoing edges and maximum internal connection as in the example where dotted circle represents the large module and colored circle represents the core.

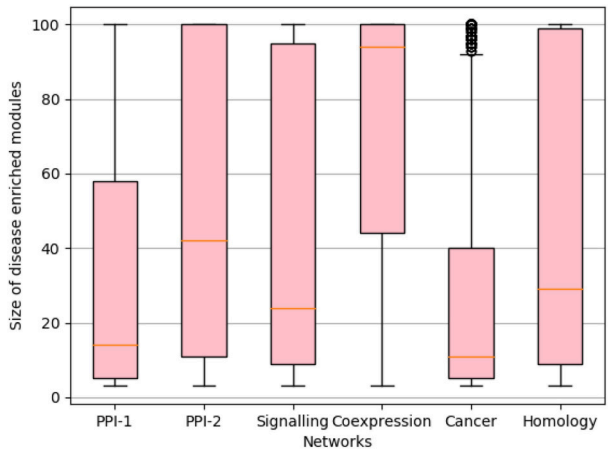


FIGURE 3 | Size distribution of disease-enriched modules identified by various methods across networks. The X-axis represent different network and Y-axis represents the size distribution of disease enriched modules. The orange line in the box-plot represents the mean of the distribution and bubbles represents the outlier data points.

few of the nodes. There is a higher chance of occurrence of multiple well-connected cores in a single large module, as depicted in **Figure 2E**. However, it is difficult to avoid merging of these modules at the time of module formation process during the modularity maximization step. Generally, modules grow quickly around a high-degree node due to the frequent merging of communities around it, whereas modules grow slowly around the section of a network having low-degree nodes. If we stop the iterative module formation early, to capture smaller communities, it often compromises on the module lying in the sparser regions of the network. Therefore, we allow the module formation step to progress until there is no change in the modularity score of the entire network. Thereafter, we perform an iterative partitioning of larger modules into multiple smaller modules. This re-clustering resulted in many smaller modules fitting the size requirements of the challenge.

2.4. Overlapping to Non-overlapping Community Assignment

For understanding the sensitivity of the overlapping and non-overlapping clustering approaches, we convert the overlapping communities to non-overlapping communities and compare their performance. Initially, we form a *base community*, which is comprised of only those nodes that got a single community assignment. To obtain non-overlapping communities the nodes that are part of multiple communities, i.e., the overlaps of the communities, needs to be assigned to one of the base community.

We have suggested the following three ways of assigning the overlap to the base community:

Random Allocation: Random allocation involves assigning the nodes in the overlap randomly to one of the base communities. The drawback of this method is that the community structure is not well-defined after the assignment.

Conductance Assignment: In order to maintain the structure of a community, the node assignment should be based on some quality measure for the community. We have used conductance. We assign nodes in the overlap to the base community with which it has the minimum conductance score.

However, while assigning a node to a base community, the conductance score for each node is independently checked against each base community. This means that, toward the end of the node assignment, the community structure need not be preserved, as all the nodes were independently assigned and the inclusion of even a single node can drastically change the community structure.

Iterative Conductance Assignment: To resolve the problem addressed in the previous approach, we follow an iterative way of assigning the node to the base community. Each node in the overlap is assigned to the base community with the minimum conductance, one after the other, and the conductance score is re-computed for the base community. This a called as a phase of community assignment.

After completion of a community assignment phase, the nodes which were part of the overlap are extracted from the base community one by one and reassigned to the community with which it has the best conductance score. This is done to avoid any bias due to the order in which the nodes were assigned. Thus, the phases are repeated iteratively till convergence, when no node changes its community. Though we do not give a proof of convergence, we have empirically observed that this approach converges after a few (typically 3–5) iterations.

2.5. Evaluating Disease Modules

The DREAM challenge organizers provided a novel framework for assessing the methods based on the number of predicted modules that are significantly associated with complex traits and diseases (with the help of GWAS data). Instead of using traditional methods that take into consideration the functional annotation or pathway databases, they used GWAS datasets. This methodology of scoring is better, unlike functional annotations that originate from a similar type of functional genomics as the networks themselves. GWAS provides an entirely orthogonal view, for validation.

2.5.1. Module Scoring Using PASCAL

PASCAL (Lamparter et al., 2016) stands for PAthway SCoring ALgorithm, which is a tool used to integrate SNP-trait associated *p*-values to incorporate gene-score and module score as illustrated in **Figure S1**. The gene score is computed by aggregating SNP-*p*-values for a GWAS dataset while correcting for confounders such as Linkage Disequilibrium (LD) correlation structure as explained in **Figure S1A**. For the module genes which are in LD and cannot be treated independently, this fast gene scoring method fuses the genes and recomputes the gene score as in **Figure S1B**. Modified Fisher method is used for computing enrichment in high scoring genes, where genes in the network become the "background gene set". The **enrichment score** is defined as the number of modules with the significant score at 5% FDR (false discovery rate) cut-off for at least one of the GWAS dataset. The final score of the method is the number of disease enriched modules it discovers.

2.6. Implementation

All the approaches in core module identification, module identification using non-overlapping community detection and overlapping to non-overlapping community assignment were implemented in Python. For modularity maximization, we have used the implementation from the NetworkX package for Python (Hagberg et al., 2008). The MCL-edge software provided by Enright et al. (2002) was used for finding clusters using MCL. The implementation of CEIL algorithm was taken from the source code provided by Sankar et al. (2015). The evaluation script was provided by DREAM challenge organizers.

3. RESULTS

Next, we study the community structure of the networks, to investigate if the disease modules are indeed clusterable, and proceed to answer the questions posed in section 1. We then show the importance of performing an overlapping community detection, and how it captures far more relevant modules. We further go on to illustrate how some knowledge of communities, in terms of "seed nodes" can positively impact the quality of clusters. Lastly, we show that overlaps of the disease modules helps in identifying comorbid diseases.

3.1. Core Module Identification Captures a Higher Number of Disease-Relevant Modules Than Traditional Community Detection Approaches

The well-known non-overlapping clustering approaches like MCL, modularity maximization and CEIL, tend to identify communities that are large and their size is dependent on the size of the network. However, disease modules are generally small. Wilber et al. (2009) have shown that small communities in these networks are biologically homogeneous. Biological homogeneity is evaluated from the functional similarity between pairs of genes, which is available from resources such as the Gene Ontology Database (Ashburner et al., 2000). They have shown that the functional similarity between pairs of genes in a small module is significantly higher than the functional similarity between all possible pairs. Core module identification methods identify smaller and structurally better communities. The size distribution of the traditional and core module based methods can be seen in the **Figure S2**.

Most of the methods that are considered have hyper-parameters; varying them could control the size and the number of modules detected. We have evaluated all the methods through an extensive grid search (parameter tuning) and report the best result for each method; the corresponding hyper-parameters are mentioned in the **Table S2**. For *MCL*, we vary the inflation (I) parameter in the range [2, 9] at intervals of 1

TABLE 2 | Results of module identification approaches on simple networks, using off-the-shelf approaches mentioned as baselines, core module identification proposed by us and Diffusion State Distance (DSD) which is the winning method of DREAM challenge.

Network	Baselines			Core module based methods				DSD
	MCL	Modularity maximization	CEIL	Perturbation	Ensemble	Recluster	Min outgoing	
(A)								
PPI 1	16 (872)	8 (262)	12 (1398)	18 (260)	16 (250)	20 (460)	22 (462)	**24 (1020)**
PPI 2	18 (1125)	9 (209)	11 (1696)	17 (284)	17 (601)	19 (1311)	**21 (1311)**	19 (445)
Signaling	9 (268)	10 (111)	6 (320)	9(180)	8 (191)	11 (144)	14 (77)	**17 (194)**
Co-expression	9 (463)	10 (194)	5 (1336)	13 (126)	12 (202)	17 (145)	20 (205)	**24 (207)**
Cancer	4 (598)	4 (164)	5 (831)	6(598)	5(249)	3(518)	**9 (114)**	7 (329)
Homology	8 (180)	6 (177)	7 (320)	7 (168)	7 (87)	7 (212)	10 (149)	**11 (212)**
Score	64	47	46	70	65	77	96	**102**
(B)								
PPI-1	0.0183	0.0305	0.0086	**0.0692**	0.064	0.0435	0.0476	0.0235
PPI-2	0.016	0.0431	0.0065	**0.0599**	0.0283	0.0176	0.0195	0.0427
Signaling	0.0336	0.0901	0.0188	0.05	0.0419	0.0764	**0.1818**	0.0876
Co-expression	0.0194	0.0515	0.0037	0.1032	0.0594	**0.1172**	0.0976	0.1159
Cancer	0.0067	0.0244	0.006	0.01	0.0201	0.0058	**0.0789**	0.0212
Homology	0.0444	0.0339	0.0219	0.0417	**0.0805**	0.033	0.0671	0.0519

The result contains, for different methods, ***(A)*** *the number of enriched modules out of the total number of predicted modules in brackets. The score in the last row represents the sum of disease modules predicted across the six different networks, and* ***(B)*** *"hit ratio", illustrating the fraction of predicted modules that are enriched.*
The numbers in bold highlight the best approach for each of the 6 networks. The results show that DSD approach predicts more enriched modules, while core-module based approaches give higher hit-ratio (ratio of enriched to total predicted modules).

and the expansion is fixed at 2. The resistance (R) parameter for *modularity* is varied in the range [0.1, 1] at intervals of 0.1. *CEIL* does not have any hyper-parameter to be tuned. **Table S3** presents the detailed results at each parameter setting. Core module identification methods are frameworks to extract compact modules and are built on top of the baseline methods. For core module identification, we have experimented with all the baseline methods and have reported the one's giving the best performance along with its hyper-parameter. The winners of DREAM challenge used Diffusion State Distance (DSD) (Cao et al., 2013, 2014) as a distance measure to perform kernel-based clustering. We have compared against their winning results. For the *perturbation* method, modularity maximization (with R as 0.1) is applied on all the perturbed networks; then consensus is taken over the modules predicted across perturbed networks. *Ensemble* uses all the baseline methods with all possible hyper-parameters. *Recluster* was done on the giant modules obtained from the best reported baseline method for that network. Therefore, the baseline method along with their hyper-parameter are reported in the table. The method for selecting nodes with *minimum outgoing edges* was applied after recluster method.

The results denote the number of enriched modules out of the predicted modules from the methods. The enrichment of a module is tested using PASCAL tool across 180 GWAS datasets. The results with best hyper-parameter setting are given in the **Table 2**; the number of disease-enriched modules identified by core module-based methods is much higher than those identified by the baseline approaches (**Table 2A**). In addition, we also show the "hit ratio" (**Table 2B**), illustrating the fraction of predicted modules that are enriched. Some methods, such as CEIL predict a large number of modules, but not many of them are enriched. On the other hand, our method, although it predicts marginally fewer modules, shows a much higher hit ratio. The reason for performance improvement on applying the proposed heuristics is due to the identification of core modules, which are smaller and structurally better, as discussed in section 2.3. From our proposed methods, min outgoing edges has the best performance with respect to number of disease-enriched modules identified, as it is a two-way refinement process (1) reclusters the giant modules obtained from baseline methods; therefore making the modules small with better internal connectivity (2) selects the nodes based on core score thus pruning away the less important nodes. The performance of our model is comparable to the winning team's performance, and in networks like PPI and Cancer, our method even outperforms the winning team's method, showing the strength of our model.

3.2. Clusterability of Disease Modules: An Analysis of Non-overlapping Community Detection

On analyzing the results of non-overlapping clustering approaches a common problem existed for all the methods, the number of enriched modules was quite less in comparison to the number of modules predicted. Also, the number of diseases enriching the modules were very less in comparison to the number of different (180) GWAS datasets available for testing.

To understand the network structure of disease module we studied the "clusterability" of disease modules. We define clusterability as the connectivity strength or the quality of the module. The ground truth disease modules are readily not available. To analyse the clusterability of a disease module, we

try to understand the connectivity between the genes that are known to have an association with the same phenotype. These disease genes need not be highly interconnected to possess the graph-theoretic community structure. This phenomenon could be explained with the help of the **Figure 4**, where, the same colored nodes represents genes associated with a disease. As is evident from the figure, there are two possibilities: (1) genes associated to disease are in the neighborhood but are not so strongly connected to qualify the definition of community, or (2) structurally well-defined community need not be associated with a particular phenotype.

To understand clusterability, we examined the cluster quality of the largest connected component (LCC) of genes (nodes) in the network associated with the same GWAS dataset. Here, the cluster quality is defined on the structural properties of the cluster. We obtain cluster quality based on modularity and conductance scores. Modularity score is the difference between the number of edges that fall within the given clusters and the expected number of edges if edges were distributed at random (Newman, 2004). Whereas, conductance is indicative of dense connections within the group, and fewer links to the rest of the network. For good quality clusters, a higher modularity score (best is 1.0) and a lower conductance (best is 0.0) are preferred.

We observed that the cluster quality of the LCC of trait-associated genes is quite poor. The cluster quality of the LCC is depicted by the heatmaps representing as shown in **Figure 5**: the X-axis represents cluster quality across 180 GWAS datasets, which are stacked one above the other in groups of 30 datasets (stacking was done to aid visualization). The Y-axis represents the six networks. All the LCC have poor modularity scores, which are close to 0 as in **Figure 5A**. The conductance score is also poor for most of the LCCs, as shown in **Figure 5B**.

Community detection methods based on optimizing cluster quality measures fail to identify disease modules because of poor modularity and conductance scores of these modules. Therefore, it is hard to identify disease modules using community detection approaches based on optimizing these cluster quality measures. However, it would be interesting to study the structurally well-defined community that could be obtained from these LCC.

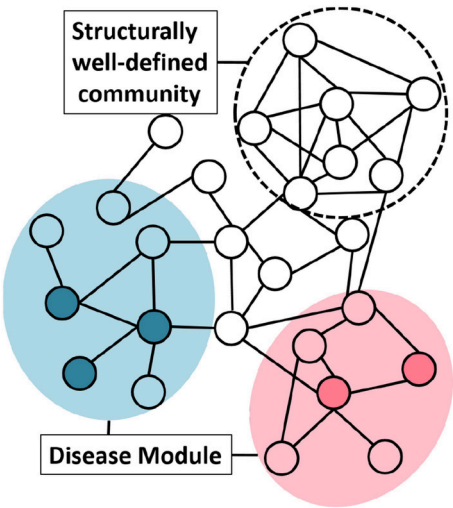


FIGURE 4 | Group of genes associated with a disease do not necessarily possess graph-theoretic community structure. Nodes with the same color represent genes associated with the same phenotype. The shaded circle over the colored nodes represents the possible disease module while a "structurally well-defined community" need not be enriched with a specific disease.

3.3. Approximating Gold Standard Disease Modules

The ground truth disease module are readily not available in order to substitute we identify structurally well-defined modules initiated from known disease-associated gene and define it as *gold-standard disease module*. We obtain the trait-associated genes from the 180 GWAS datasets and call them as *disease seed nodes*. We explicitly try to enforce the community structure into these groups by adding the neighborhood nodes using the seed expansion process.

3.3.1. Gold Standard Modules Exhibit Clusterability

The modules obtained after this disease seed node expansion procedure were checked for enrichment using the PASCAL tool as described in section 2.5.1. The enriched communities so obtained are called as the *gold-standard disease modules*. These modules have proper community structure and are curated from the significant disease nodes. The statistics pertaining to the number of disease seed nodes obtained and the number of gold-standard modules identified are shown in **Table 3**. The percentage of the seed nodes covered in the enriched communities represents that these seed nodes have a well-structured disease neighborhood around them. "Disease spread" represents the number of GWAS datasets from out of 180 of them that could be identified in a particular network. The empirical results as in **Table 3** suggest that many diseases have a good community structure in the PPI-1 network. These results also show that prior knowledge of disease seed nodes improves the performance of community identification by ten times as opposed to purely network driven community detection (**Table 2**). For example, in the case of PPI-1, we find 337 disease-enriched modules with this approach, compared to 22 from section 2.3.

The disease seed node expansion procedure is helpful in identifying many disease enriched modules in comparison to the methods described in section 2.3. However, along with the increase in the number of enriched modules, there is also an increase in the number of non-enriched modules. We now study the difference in cluster quality of the enriched and the non-enriched modules.

We calculate the cluster quality of all the modules, predicted by the gold standard module identification process, using modularity and conductance scores. The predicted modules are divided into two sets—enriched and non-enriched modules—based on the enrichment predicted by the PASCAL tool (section 2.5.1). The distributions of cluster quality scores for the enriched and non-enriched modules across the six networks were then compared using notched box plots.

The distributions of the modularity and conductance scores of enriched and non-enriched modules can be visualized in the **Figure 6**. The notch represents the confidence interval around the median. The visual interpretation of these notches is that,

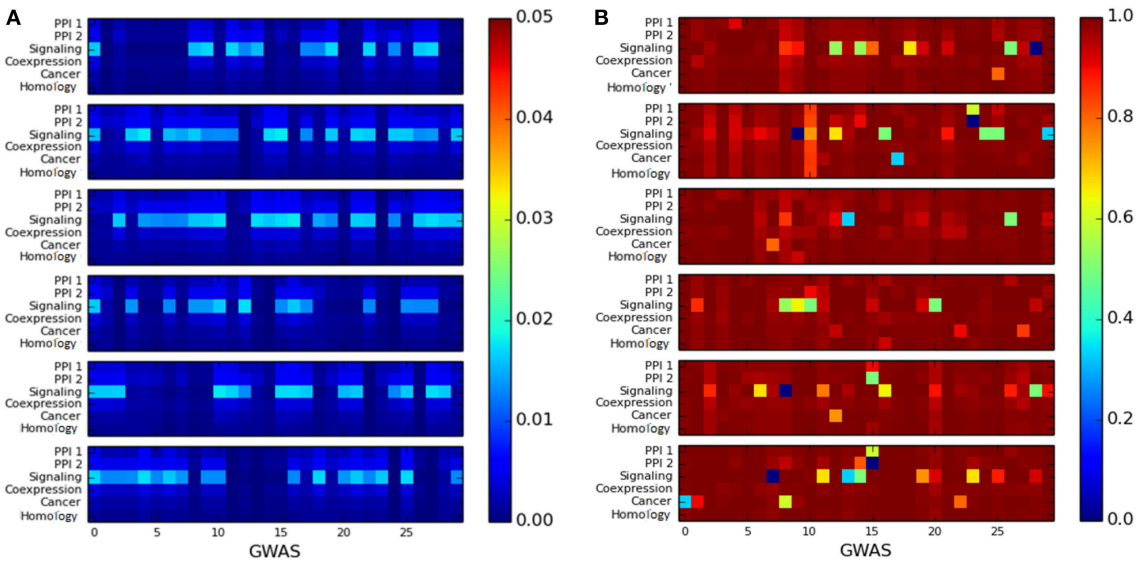


FIGURE 5 | Heatmap for measuring the quality of the largest connected component of genes associated with a single disease using: **(A)** modularity score, where higher score is preferable and **(B)** conductance score, where lower score is preferable. The X-axis represents the 180 GWAS datasets and Y-axis represents the six networks. The color bar on the right represents the color coding for values marked in the heat map.

if notches of box plot of two distributions do not overlap, then their medians differ with 95% confidence. The mean of the distributions of scores for enriched and non-enriched counterparts vary significantly as there is no overlap between the notches of the two distributions. This variation is quite significant for PPI-2 and signaling networks.

We can conclude from the study that disease-enriched modules in all the networks have better clusterability properties. So the predicted modules could be ranked on the basis of modularity or conductance scores, and the higher ranked modules are more likely to be disease modules.

3.3.2. Amount of Disease Seed Nodes Required for Expansion

We took 10^{-4} and 10^{-6} as p-value thresholds, to identify disease seed nodes. The identified disease seed nodes across the set of 180 GWAS dataset is quite large in comparison to the total number of genes in the network as can be seen in **Table 3**. For example, in the case of PPI-1, 5436 disease seed nodes are identified, whereas there were 17397 genes in the network (from section 2.1), which means 30% of the network is a part of disease seed nodes.

The percentage of genes covered in the disease modules indicate that not all disease seed nodes are required for disease module identification. Here, we proceeded to analyse the amount of known disease seed nodes required for expansion, and how the performance is impacted knowing a fewer number of disease seed nodes. We randomly selected 10, 50, and 80% of the known seed nodes (with a p-value cutoff of 10^{-6}) and performed disease seed node expansion from these and observed the enriched modules obtained. This step of randomly selecting k% nodes was repeated five times to avoid any bias due to a single run. The enriched modules reported in **Table 4** shows the average of the modules predicted in five runs. It is observed that for some networks like PPI-1, PPI-2, and signaling, increasing the number of known seed nodes improves the number of disease modules recovered. In other networks, namely, homology, cancer and co-expression, the number of known seed nodes did not substantially change the number of disease modules identified.

3.4. Disease Modules Are Naturally Overlapping and Transcription Factors Mostly Lie in the Overlaps of Disease Modules

From the gold standard module identification procedure we obtain overlapping communities and the drastic increase of almost 10 times in the number disease modules identified in comparison to non-overlapping methods (**Tables 2**, **3**) suggests that "overlapping methods" should be a preferred choice for disease module identification. We also try to find out the biological relevance for the nodes that are part of multiple communities. An *overlap* is defined as the nodes that are shared by a pair of overlapping modules. We find that nodes that lie in the overlap of the gold standard disease modules are mostly transcription factors (TF). **Figure 7** shows the box plot of the distribution of the number of enriched modules the TFs are part of. Transcription factors regulate the expression of multiple genes and hence affect multiple pathways of varying functions. Since TFs control different functions, they are expected to be found in overlapping regions of the disease modules.

3.5. Overlapping Community Detection in the Absence of Known Disease Seed Nodes

As we established the importance of overlapping community detection for disease module identification it is also necessary to

TABLE 3 | Gold standard disease modules identified from disease seed node expansion, keeping *p*-value threshold as 10^{-4} and 10^{-6} across 180 GWAS datasets.

Network	Significance threshold 10^{-4}				Significance threshold 10^{-6}			
	# Seed nodes	Enriched	Disease spread	Seed nodes in enriched (%)	# Seed nodes	Enriched	Disease spread	Seed nodes in enriched (%)
PPI 1	5436	337 (5433)	52	39.09	3103	266 (3101)	57	37.12
PPI 2	3876	130 (3844)	28	21.05	2267	103 (2250)	32	21.53
Signaling	1893	158 (1840)	36	31.80	1174	126 (1139)	44	37.39
Co-expression	4099	174 (4094)	34	53.86	2406	152 (2404)	38	54.61
Cancer	4507	6 (4429)	5	2.37	2555	2 (2522)	2	1.76
Homology	3227	28 (3154)	7	13.39	1861	14 (1826)	6	10.31

The number of seed nodes obtained is mentioned in the 2nd and 6th column. The 3rd and 7th columns show the number of enriched modules against the predicted ones mentioned in the brackets. Disease spread represents the number of unique GWAS datasets identified across all the predicted modules. Percentage of seed nodes covered in the module is also tabulated.

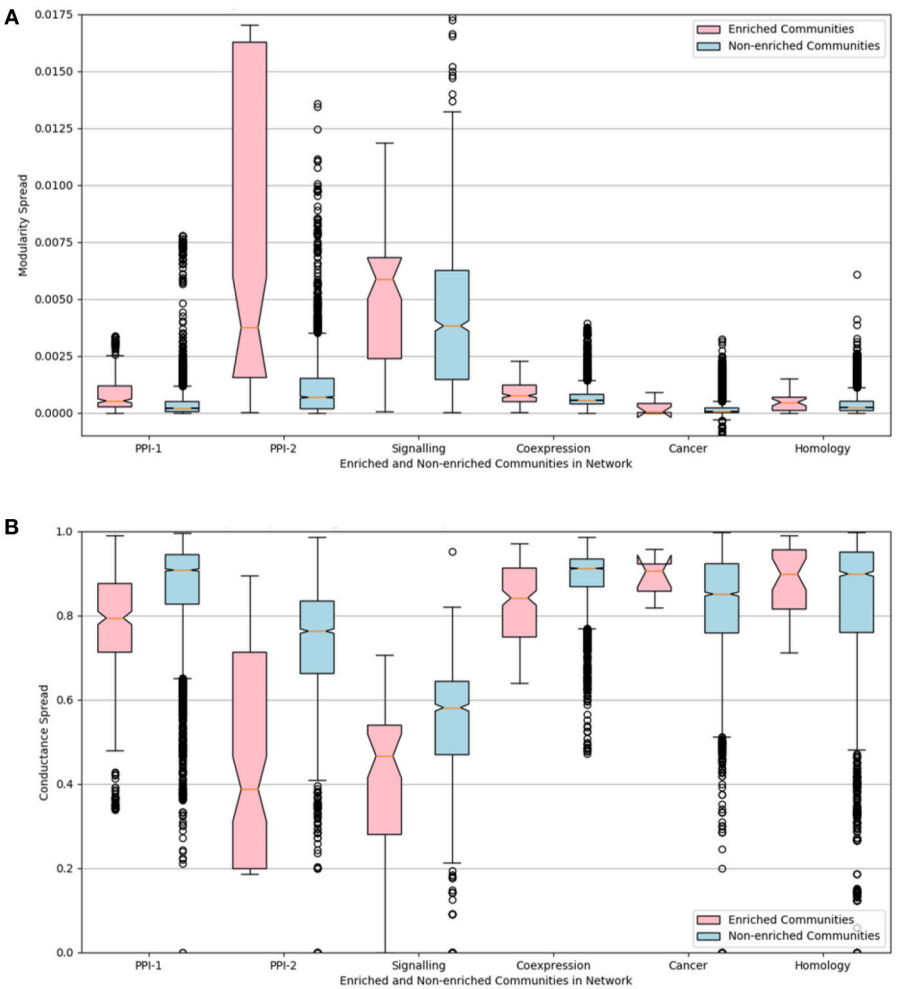


FIGURE 6 | Notched box-plot representing **(A)** modularity and **(B)** conductance of enriched modules (pink) compared to non-enriched modules (blue) across the six networks. High modularity and low conductance is preferred for better quality clusters. Owing to the lack of overlap between the notches of the two distribution, the enriched modules have a significantly higher modularity and lower conductance score in comparison to non-enriched modules. **A notched box plot** is a graphical way of representing data. The box represents the interquartile range (IQR) of the data, where 50% of the data fall. The middle line denotes the median of the data. The top whisker is 1.5 times more than 75 percentile (Q3), and bottom whisker is 1.5 times lesser than 25 percentile. The notch represents the confidence interval around the median. The visual interpretation of these notches is that, if notches of box plot of two distributions do not overlap, then their medians differ with 95% confidence.

TABLE 4 | The number of modules predicted when starting with different fraction of initial known seed nodes (10^{-6}).

Disease Seed	PPI 1	PPI 2	Signaling	Co-expression	Cancer	Homology
100% 10^{-6}	266 (3101)	103 (2250)	126 (1139)	152 (2404)	2 (2522)	14 (1826)
80% 10^{-6}	165.8 (2147.6)	90 (1357.2)	53.4 (597.6)	152.2 (1859.4)	5 (1484.2)	14 (1221.8)
50% 10^{-6}	111 (1378.4)	60.6 (897.4)	35.2 (407.6)	105.8 (1167.4)	5.2 (979.6)	13 (793.8)
10% 10^{-6}	40 (292.8)	20.6 (202.6)	16.8 (98.2)	29 (236)	8.4 (216.8)	12 (174.2)

Here, the value represents the number of enriched modules out of the total number of modules predicted mentioned in brackets.

explore this approach in the absence of known disease seed nodes. We have selected HITS and spread hubs, which selects nodes based on their degree, to identify seed nodes. For each network, we kept the number of seed nodes fixed as for gold-standard disease module, and selected those many seed nodes using HITS and spread hub. The enriched and predicted modules obtained after seed node expansion can be found in the **Table S4**. This table also compares against the gold standard disease module results. It is observed that the best results in terms of the number of enriched modules predicted, is obtained for the PPI-1 network.

The performance of unsupervised seed node expansion is visually compared with the gold standard module identification process with the help of scatter-plots as in **Figure 8**. The X-axis and the Y-axis of the plot represent the number of enriched modules as predicted by gold-standard module identification and unsupervised seed node expansion respectively. Different colors in the plots correspond to different networks as mentioned in the legend. For the same network, the plot shows multiple bubbles; those are with respect to the different number of seed nodes used for expansion. The line in the plot is for $x = y$, where the performance of disease seed node expansion is similar to unsupervised seed node expansion. As is quite intuitive, all the bubbles are below the partition line which means that the performance of disease seed node expansion is consistently better. It is observed that the performance of unsupervised seed node expansion on PPI-2 is comparable to its gold-standard disease module counterpart. Also, **Figure 8B** shows that the performance of spread hub as seed node selection is quite close to the disease seed node expansion as all the bubbles are much closer to the partition line.

3.5.1. Sensitivity Analysis of Non-overlapping and Overlapping Clustering Approaches

The methods based on optimizing a "quality function," such as conductance or modularity, non-overlapping communities, which means a node is part of a single module. The other class of methods we concern ourselves with are the non-overlapping clustering approaches using seed node based expansion methods. Below, we perform a detailed comparison of both classes of methods. We compare the number of enriched modules obtained from seed node based expansion methods with the quality function based approaches to understand the superiority of one over the other.

However, for a fair comparison between the approaches all of them should have a similar setup that is they should have either overlapping or non-overlapping communities.

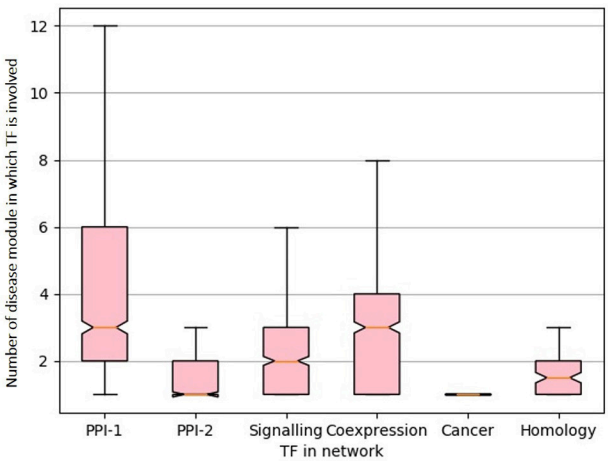


FIGURE 7 | Box-plot representing transcription factors involved in multiple diseases. The X-axis represent different network and Y-axis represents distribution of number of disease module a TF is involved. The notches in the box-plot represents the mean of the distribution.

Therefore, we convert overlapping communities to non-overlapping communities using methods defined in section 2.4

The performance of the quality function-based community detection and seed node expansion methods are compared in the **Table 5**. The results suggest that identifying important nodes in the network and localizing communities around them is a better way of performing disease module identification, where compared to growing and merging communities from all possible nodes as done in quality function based approaches.

3.6. Overlapping Disease Modules Helps in Identifying Comorbid Diseases

We proceeded to derive useful insights from gold standard modules by studying comorbidity among diseases, i.e., those disease which have chances of co-occurring together. Comorbidity study is done by identifying diseases associated with the same disease enriched modules. **Figure 9** shows a box-plot, representative of the distribution of the number of diseases that are associated with an enriched module; these modules are identified by various approaches which are already discussed in this and previous chapters. The distribution shows that there are multiple diseases associated with a disease enriched modules,

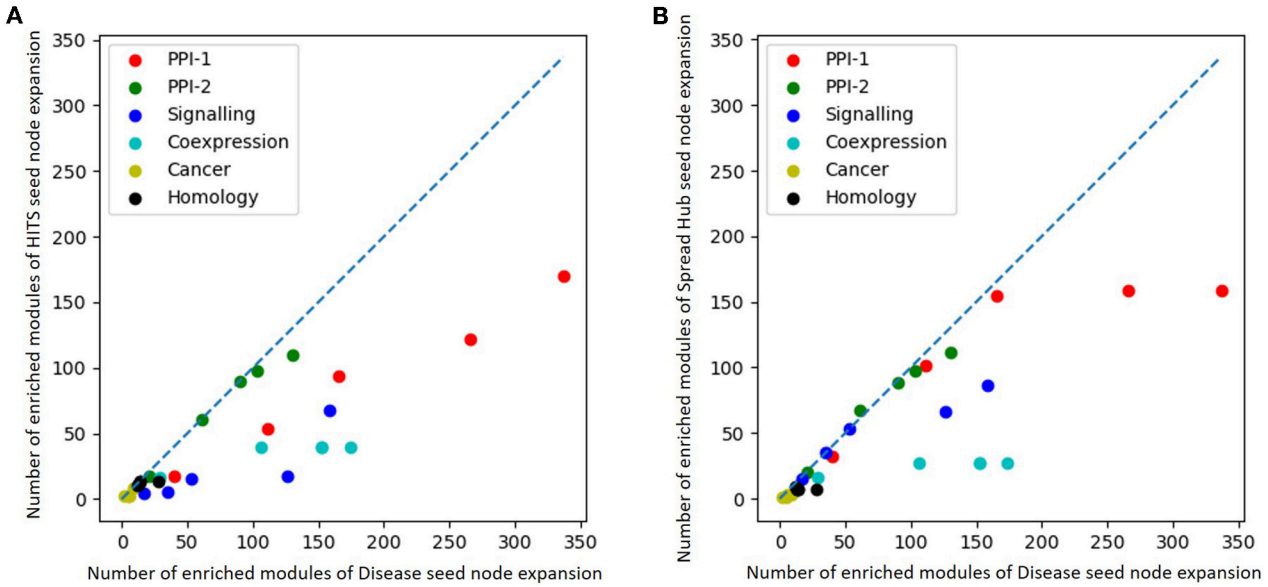


FIGURE 8 | Scatter plot comparing number of enriched module predicted by **(A)** HITS **(B)** Spread hub seed node expansion against disease seed node expansion. The X and the Y axis represents the number of enriched modules disease seed node and unsupervised seed node expansion respectively. Different colored bubbles represents different network as mentioned in the legend. The partition line in the plot is x=y where performance of disease seed node expansion is similar to unsupervised seed node expansion.

TABLE 5 | Comparing quality function-based community detection with seed nodes expansion method after converting overlapping to non-overlapping communities for fair comparison.

Network	Quality function based		Seed expansion based		
	Modularity	CEIL	Disease	HITS	Spread hub
PPI 1	8 (262)	12 (1398)	26 (283)	24 (921)	28 (888)
PPI 2	9 (209)	11 (1696)	16 (191)	11 (659)	16 (207)
Signaling	10 (111)	6 (320)	15 (219)	10 (192)	12 (183)
Co-expression	10 (194)	5 (1336)	24 (234)	15 (743)	11 (240)
Cancer	4 (164)	5 (831)	11 (209)	6 (239)	7 (962)
Homology	6 (177)	7 (320)	9 (159)	3 (172)	9 (185)
Total	47	46	101	69	83

The values in the table represent the number of enriched modules along with the predicted modules in the bracket.

especially in PPI-1. Modules enriched for multiple diseases are helpful in finding the association between the diseases. A module represents a group of genes responsible for diseases. Thus, if a person gets a particular disease due to improper functioning of few genes, then (s)he is likely to get another disease whose underlying responsible genes are the same. This study can help in answering questions such as, if a person has a particular disease, then how likely he can have another disease. As PPI-1 has the highest number of comorbid associations, we choose modules identified on this network for co-morbidity study.

We formed a comorbid network where the nodes are different diseases as shown in **Table S1** and the edges are indicative of a module being enriched with the connected disease nodes. We consider the modules that are getting enriched with multiple diseases and connect all these diseases with an edge, and we also keep an edge count as to how many times those two diseases occurred together. **Figure 10** shows the comorbid network created from the association between diseases of enriched modules on PPI-1, here we have kept top 50% associations based on the edge count.

The top associations represent the most frequently co-occurring diseases identified, based on the modules enriched with multiple diseases. Higher association between two disease means that there is more evidence for their correlation, as they have been grouped together more number of times. From **Figure 10**, anthropomorphic disease are seen to be connected with most of the other diseases suggesting that it is linked with many of the diseases. Further, the links between Glucose Metabolism and Lipid/Heart are also not very surprising, given the remarkable co-ocurrences of diabetes, coronary heart diseases, and hypercholesterolemia etc.

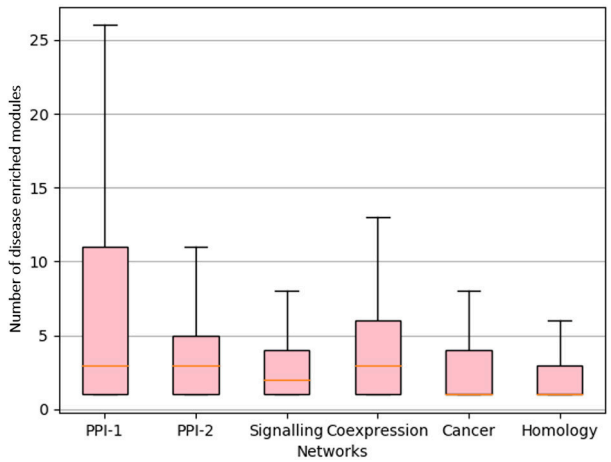


FIGURE 9 | Distribution of number of GWAS datasets associated with the disease enriched module identified by various approaches. The X-axis represents different networks and Y-axis represents number of GWAS datasets associated with the disease module. The orange line represents the mean of the distribution.

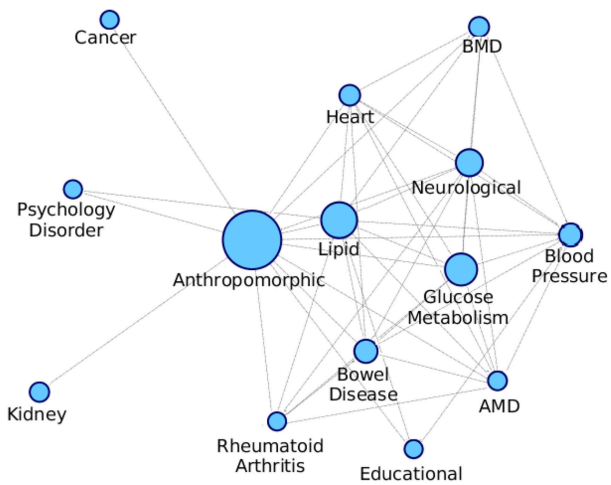


FIGURE 10 | Comorbidity network showing associations between different diseases, created on the basis of number of GWAS datasets associated with enriched modules. The nodes are sized based on the number of GWAS datasets associated with the disease. AMD stands for Advanced Macular Disorder and BMD stands for Bone Mineral Density. These diseases are the same as those in **Table S1**. Hepatitis-C mentioned in the table is not in the comorbid network as there is minimal evidence of it being associated with any other disease.

4. DISCUSSION

The identification of communities in networks is a well-studied problem in computer science. In this DREAM challenge, the goal was to identify such modules, or communities, in various biological networks, and study their association with diseases. In the present study, we examined various approaches for community detection, and their applicability to biological networks, to identify disease-relevant modules. Notably, we illustrate the importance of identifying smaller "core" communities compared to standard non-overlapping clustering algorithms. Further, we analyse the need and importance of overlapping communities and the utility of seed nodes or partial knowledge in greatly improving the prediction of biological relevant disease modules from diverse networks.

We have three key results. First, we show that well-known non-overlapping clustering approaches fail to identify sufficient number of relevant disease modules. Our core-module based identification methods, which identify smaller and structurally better communities, could identify larger number of disease-enriched modules than those identified by well-known non-overlapping community detection approaches. The state-of-the-art non-overlapping clustering approaches detect large communities and the core module identification approaches detect small communities as can be seen in the **Figure S2**. In almost all the cases, we saw an improvement in the performance on downsizing the size of the communities. It is important to note that this was also affected by the DREAM challenge evaluation, which mandated the identification of smaller communities ranging from 3–100 nodes. Nevertheless, such smaller communities are more common in biological networks (Wilber et al., 2009), and can indeed capture more disease-relevant communities as observed in the results. Another important observation was that the different networks provided in the challenge present diverse views of the interactions happening in the cell. Therefore, each network has different network properties and consequently need different approaches to identify disease modules in them. For example, PPI-1 had smaller-sized disease modules than PPI-2 as can be observed in **Figure 9**, and hence *Multiple Core Identification* was able to perform better than *MCL* as the former method downsizes the size of the community. *Min Outgoing edges* further reduced the size of the module thus improving on the number of modules identified. For signaling and co-expression network we know that genes interacting with more number of other genes are biologically more active (Vidal et al., 2011) and this could be seen in the results – method *min outgoing edges* which gives more importance to nodes with higher degree showed a remarkable improvement over *modularity maximization* method. It was hard to achieve good performance for cancer network as despite using known disease nodes in **Table 3**, module identification gave very poor number of disease enriched modules. For the homology network, ensemble with *min outgoing edges* was useful. Overall, the number of disease-enriched modules identified by core module-based methods was higher than those identified by the baseline approaches. The reason for performance improvement on applying the proposed heuristics is due to the identification of core modules, which are smaller and structurally more informative.

Second, we investigated the clusterability properties of disease modules and illustrated that there do exist well-defined communities, but a overlapping clustering approach was important to capture them, particularly in face of the fact that most proteins have multiple functionalities or cause different diseases. Owing to the lack of ground truth disease module, the enriched modules identified after exploiting domain

knowledge—the "gold standard modules"—have better cluster quality than their non-enriched counterparts. This indicates that disease modules, when carefully identified with the help of some known disease nodes, possess good clusterability.

Third, we showed that information on "seed nodes" underlying these modules can substantially improve the identification of disease-relevant modules. As the fraction of disease seed node increases number of identified disease enriched modules increases (**Table 4**). Interestingly when disease nodes are not known identifying seed nodes using spread-hubs and doing seed expansion on it perform equally well especially for lower fraction of seed nodes as can be seen in **Table S4**. Thus further supporting the fact, overlapping community detection is a better way to identify disease modules. Also, the overlaps between gold standard module identification are also useful for identifying co-occurring diseases, such that occurrence of one disease results is a signal that the other one can also occur. We also show that localizing community discovery around a network-centric, biologically relevant node (seed node) offers a clear advantage for disease module identification in comparison with a completely unsupervised approach. Domain guidance is essential and should be leveraged upon whenever possible. We observe this when one compares the performance of quality function based methods with the seed expansion strategy than extant approaches as in **Table 5**.

Our study underlines the need to develop biologically motivated clustering algorithms that are able to better capture "disease community structure" and notably, emphasizes the importance of overlapping clustering approaches to reliably identify disease-relevant modules and comorbidity networks from diverse biological datasets. Notably, our results underline the importance of overlapping community detection and makes the case for further investigation into such methods, rather than non-overlapping community identification, in the case of biological networks.

DATA AVAILABILITY

The datasets and codes for this study are available on GitHub: https://github.com/RamanLab/DiseaseModuleIdentification.

AUTHOR CONTRIBUTIONS

BT, KR, and BR designed the experiments. BT performed the experiments. BT, SP, KR, BR, and HS analyzed the results and wrote the manuscript.

FUNDING

This work was supported partly by a grant under the VAJRA faculty scheme (grant number: VJR/2017/000187) of the Govt. of India to SP, a research grant from Intel India to BR and New Faculty Initiation Grant from Industrial Consultancy & Sponsored Research, Indian Institute of Technology Madras (grant number: BIO/16-17/856/NFIG/HIMA) to HS.

ACKNOWLEDGMENTS

We would like to thank Karthik Azhagesan for working as a team in DREAM challenge and helping with the initial experiments. We would like to acknowledge Manikandan Narayanan, Ujjawal Soni, and Akshat Sinha for various fruitful discussions. We also acknowledge Disease Module Identification DREAM challenge organizers for providing us the data used in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00164/full#supplementary-material

REFERENCES

Andersen, R., Chung, F., and Lang, K. (2006). "Local graph partitioning using pagerank vectors," in *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)* (Berkeley, CA: IEEE), 475–486.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29. doi: 10.1038/75556

Asur, S., Ucar, D., and Parthasarathy, S. (2007). An ensemble framework for clustering protein protein interaction networks. *Bioinformatics* 23, i29–i40. doi: 10.1093/bioinformatics/btm212

Bader, G. D., and Hogue, C. W. (2002). Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* 20:991. doi: 10.1038/nbt1002-991

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512. doi: 10.1126/science.286.5439.509

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12:56. doi: 10.1038/nrg2918

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2010). Ncbi geo: archive for functional genomics data sets 10 years on. *Nucleic Acids Res.* 39(Suppl_1) D1005–D1010. doi: 10.1093/nar/gkq1184

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., et al. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* 30, i219–i227. doi: 10.1093/bioinformatics/btu263

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., et al. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE* 8:e76339. doi: 10.1371/journal.pone.0076339

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Lamparter, D., Lin, J., et al. (2018). Open community challenge reveals molecular network modules with key roles in diseases. *bioRxiv*. doi: 10.1101/265553

Cowley, G. S., Weir, B. A., Vazquez, F., Tamayo, P., Scott, J. A., Rusin, S., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* 1:140035. doi: 10.1038/sdata.2014.35

Dongen, S. V. (2000). *Performance Criteria for Graph Clustering and Markov Cluster Experiments*. Technical report, National Research Institute for Mathematics and Computer Science.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Vol. 1. New York, NY: Springer.

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Ghosh, J., and Acharya, A. (2013). "Cluster ensembles: Theory and applications," in *Data Clustering: Algorithms and Applications* (New York, NY), 551–570.

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc. Natl Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Hagberg, A., Swart, P., and S Chult, D. (2008). *Exploring Network Structure, Dynamics, and Function Using NetworkX*. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402:C47.

Jaccard, P. (1901). Etude de la distribution florale dans une portion des alpes et du jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.

Ji, J., Zhang, A., Liu, C., Quan, X., and Liu, Z. (2014). Survey: Functional module detection from protein-protein interaction networks. *IEEE Trans. Knowl. Data Eng.* 26, 261–277. doi: 10.1109/TKDE.2012.225

Kannan, R., Vempala, S., and Vetta, A. (2004). On clusterings: good, bad and spectral. *J. ACM* 51, 497–515. doi: 10.1109/SFCS.2000.892125

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS Comput. Biol.* 12:e1004714. doi: 10.1371/journal.pcbi.1004714

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* 14:61. doi: 10.1038/nmeth.4083

Li, Y., Calvo, S. E., Gutman, R., Liu, J. S., and Mootha, V. K. (2014). Expansion of biological pathways based on evolutionary inference. *Cell* 158, 213–225. doi: 10.1016/j.cell.2014.05.034

Newman, M. E. (2004). Analysis of weighted networks. *Phys. Rev. E* 70:056131. doi: 10.1103/PhysRevE.70.056131

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Sankar, M. V., Ravindran, B., and Shivashankar, S. (2015). "Ceil: A scalable, resolution limit free approach for detecting communities in large networks," in *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires: Association for the Advancement of Artificial Intelligence Press), 2097–2103.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, Vol 39. New York, NY: Cambridge University Press.

Singh, R., Xu, J., and Berger, B. (2006). "Struct2net: integrating structure into protein-protein interaction prediction," in *Biocomputing 2006* (Wailea, HI: World Scientific), 403–414.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13:966. doi: 10.1038/nmeth.4077

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Whang, J. J., Gleich, D. F., and Dhillon, I. S. (2016). Overlapping community detection using neighborhood-inflated seed expansion. *IEEE Trans. Knowl. Data Eng.* 28, 1272–1284.doi: 10.1109/TKDE.2016.2518687

Wilber, A. W., Doye, J. P., Louis, A. A., and Lewis, A. C. (2009). Monodisperse self-assembly in a model with protein-like interactions. *J. Chem Phys.* 131:175102. doi: 10.1063/1.3243581

Zanzoni, A., Soler-López, M., and Aloy, P. (2009). A network medicine approach to human disease. *FEBS Lett.* 583, 1759–1765. doi: 10.1016/j.febslet.2009.03.001

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.

Copyright © 2019 Tripathi, Parthasarathy, Sinha, Raman and Ravindran. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 24 July 2019
doi: 10.3389/fgene.2019.00607

QS-Net: Reconstructing Phylogenetic Networks Based on Quartet and Sextet

Ming Tan[1†], Haixia Long[2†], Bo Liao[1,2], Zhi Cao[1*], Dawei Yuan[3], Geng Tian[3], Jujuan Zhuang[4] and Jialiang Yang[2,5*]*

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, [2] School of Information Science and Technology , Hainan Normal University, Haikou, China, [3] Geneis (Beijing) Co. Ltd., Beijing, China, [4]Department of Mathematics, Dalian Martine University, Dalian, China, [5] Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

OPEN ACCESS

Edited by:
Marco Antoniotti,
University of Milano-Bicocca, Italy

Reviewed by:
Gianluca Della Vedova,
University of Milano-Bicocca, Italy
Mohammed El-Kebir,
University of Illinois at Urbana-Champaign, United States

***Correspondence:**
Bo Liao
dragonbw@163.com
Zhi Cao
66384436@qq.com
Jialiang Yang
jialiang.yang@mssm.edu

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 24 September 2018
Accepted: 11 June 2019
Published: 24 July 2019

Citation:
Tan M, Long H, Liao B, Cao Z, Yuan D, Tian G, Zhuang J and Yang J (2019) QS-Net: Reconstructing Phylogenetic Networks Based on Quartet and Sextet. Front. Genet. 10:607. doi: 10.3389/fgene.2019.00607

Phylogenetic networks are used to estimate evolutionary relationships among biological entities or taxa involving reticulate events such as horizontal gene transfer, hybridization, recombination, and reassortment. In the past decade, many phylogenetic tree and network reconstruction methods have been proposed. Despite that they are highly accurate in reconstructing simple to moderate complex reticulate events, the performance decreases when several reticulate events are present simultaneously. In this paper, we proposed QS-Net, a phylogenetic network reconstruction method taking advantage of information on the relationship among six taxa. To evaluate the performance of QS-Net, we conducted experiments on three artificial sequence data simulated from an evolutionary tree, an evolutionary network involving three reticulate events, and a complex evolutionary network involving five reticulate events. Comparison with popular phylogenetic methods including Neighbor-Joining, Split-Decomposition, Neighbor-Net, and Quartet-Net suggests that QS-Net is comparable with other methods in reconstructing tree-like evolutionary histories, while it outperforms them in reconstructing reticulate events. In addition, we also applied QS-Net in real data including a bacterial taxonomy data consisting of 36 bacterial species and the whole genome sequences of 22 H7N9 influenza A viruses. The results indicate that QS-Net is capable of inferring commonly believed bacterial taxonomy and influenza evolution as well as identifying novel reticulate events. The software QS-Net is publically available at https://github.com/Tmyiri/QS-Net.

Keywords: phylogenetic network, reticulate evolution, sextet, bacterial taxonomy, influenza reassortment

INTRODUCTION

Phylogenetic tree is usually utilized to show the evolutionary history of a set of biological entities or taxa. However, the tree-like topology cannot represent reticulate evolutionary events, such as horizontal gene transfer (HGT), hybridization, recombination, or reassortment, which have been shown to be critical in genotypic diversity, related phenotypes, estimations of evolutionary history, and virus emergence and immune evasion (Fenderson and Bruce, 2008; Vijaykrishna et al., 2015; Bastide et al., 2018). For example, HGT, also known as lateral gene transfer (LGT), promotes the diversification of microorganisms on the evolutionary time scale. This mechanism

can change the types and characteristics of bacteria and plays a major role in the genetic diversity of bacteria (Ochman et al., 2000). In the long run, it may be the dominant force affecting genes in most prokaryotes. Recombination is a major source of genotypic diversity and a core force for the formation of genome and related phenotypes (Leducq et al., 2017). Reassortment is responsible for most antigenic shifts of influenza virus (Nelson et al., 2008). Hybridization has been shown to be the main evolutionary mechanism for plants and some animals (Rieseberg et al., 2000; Yu et al., 2011).

A phylogenetic network can serve as an alternative to phylogenetic tree. When the evolutionary history of a sequence set contains reticulate events (Huson et al., 2010), generally speaking, phylogenetic networks can be divided into explicit and implicit networks. The implicit phylogenetic networks, such as split network, are often adopted to illustrate incompatible data and capture conflicting signals in a data set. With the increasing sequencing data, phylogenetic networks have become more and more important in molecular evolution.

Over the past decades, many methods have been proposed for reconstructing phylogenetic trees or networks. The most common type of method reconstructs a network directly from the original character data, usually through a parsimony or maximum-likelihood criterion. Methods in this category include Spectronet (Huber et al., 2002), maximum pseudo-likelihood estimation (Yu and Nakhleh, 2015), HGT maximum parsimony (Park et al., 2010), PhyloNetwork (Solís-Lemus et al., 2017), inferring phylogenetic networks using PhyloNet (Wen et al., 2018), and SNaQ (Claudia and Cécile, 2016). However, these methods are inefficient computationally and tend to overestimate the actual number of reticulate events in the evolutionary history (Huelsenbeck, 1995; Park et al., 2010). The second widely used method is the distance-based method, which first builds a genetic distance matrix for a taxa set and then reconstructs the phylogenetic network from the distance matrix. Methods in this category include Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), FastME (Lefort et al., 2015), ASTRID (Vachaspati and Warnow, 2015), tree-average distances method (Willson, 2013), and large-scale Neighbor-Joining with NINJA (Wheeler, 2009). The distance-based methods are very fast compared with character-based methods, but they have a disadvantage in terms of reconstruction accuracy. The third kind of methods reconstructs phylogenetic networks from weighted triplets and quartets because they can retain more information than distances. Methods in this category include local maximum likelihood using triplets (Ranwez and Gascuel, 2002), Quartet-Net (Yang et al., 2013), tree with strong combinatorial evidence (Berry and Gascuel, 2000), QNet (Grünewald et al., 2007), SuperQ (Grunewald et al., 2013), DistiQue (Sayyari and Mirarab, 2016), level 1 network from a dense quartet (Keijsper and Pendavingh, 2014), and weighted QMC (Avni et al., 2015). In addition, there are other methods using statistical models such as stochastic local search method (Tria et al., 2010), clusters (Van Iersel et al., 2010), Bayesian inference (Zhang et al., 2017), statistical model (Pickrell and Pritchard, 2012), and Monte Carlo method (Eslahchi et al., 2010).

Quartet-Net (Yang et al., 2013) is a method for reconstructing phylogenetic networks from a set of weighted triplets and quartets, which uses parsimony information sites to calculate triplet and quartet weights directly from multiple sequence alignment (MSA). Based on the calculated triplet and quartet weights, Quartet-Net then performs a split expanding process to obtain all full splits and their weights, which will transform to an evolutionary tree or network. The method is a generalization of Split-Decomposition (Bandelt and Dress, 1992). In this paper, we further generalize Quartet-Net and propose a novel method called QS-Net to reconstruct evolutionary networks based on weighted quartets and sextets. The analysis of artificial and real data sets shows that this method can reconstruct a more accurate phylogeny when the sequence data are generated from complicated evolutionary scenarios involving many reticulate events and identifies novel reticulate evolution and reassortment events.

MATERIALS AND METHODS

Background: Split and Split Weight

For a taxa set $S = \{S_1, S_2, \ldots, S_n\}$ of size n, a split consisting of two disjoint non-empty subsets of S is denoted by A | B that is, A and B. If A and B contain all the taxa in S, then A | B is called a full split; otherwise, it is called a partial split. In a phylogenetic tree, each edge is a full split that divides the tree into two parts, while in a phylogenetic network, a group of parallel edges with equal length represents a full split. If |A| = 1 or|B| = 1, the split A|B is called a trivial split. For example, the phylogenetic tree in **Figure 1A** contains five trivial full splits, such as a|bcdef, and three non-trivial full splits de|abcf, bc|adef, and ade|bcf. In general, a split A|B with |A| = m and |B| = n is called an m|n split. In addition, W(A|B) represents the evolutionary distance between taxa groups A and B. If A or B contains more than two taxa, then W(A|B) calculates the distance between the common ancestor of A and B. For example, W(a|de) = 2, W(d|ae) = 1 in **Figure 1A**, W(a|d) represents the evolutionary distance between taxa a and d, and therefore, through **Figure 1A** and these definitions, we can get this equation W(a|d) = W(a|de) + W(ae|d).

For an MSA, a simple parsimony-based method is used to estimate the weights of quartets and sextets. For example, if the character in a site is the same for taxa a, b, and c and for taxa d, e, and f, but different for a and d, then the site is defined to support the split abc | def. For any sextet abc | def, its weight W(abc|def) is defined to be the proportion of total number of sites supporting it in the MSA. The weight of a quartet say ab|cd is calculated in a similar way. After all the quartet and sextet weights are obtained, an ever-expanding process is performed based on these weights to all full splits and their weights. As shown in previous literatures (Bandelt and Dress, 1992; Yang et al., 2013), reconstructing a phylogenetic tree or network is equivalent to calculating all the full splits and their weights. Thus, we have obtained the reconstructed tree or network by this process, which could be shown by a software SplitsTree4 (Huson and Bryant, 2006).

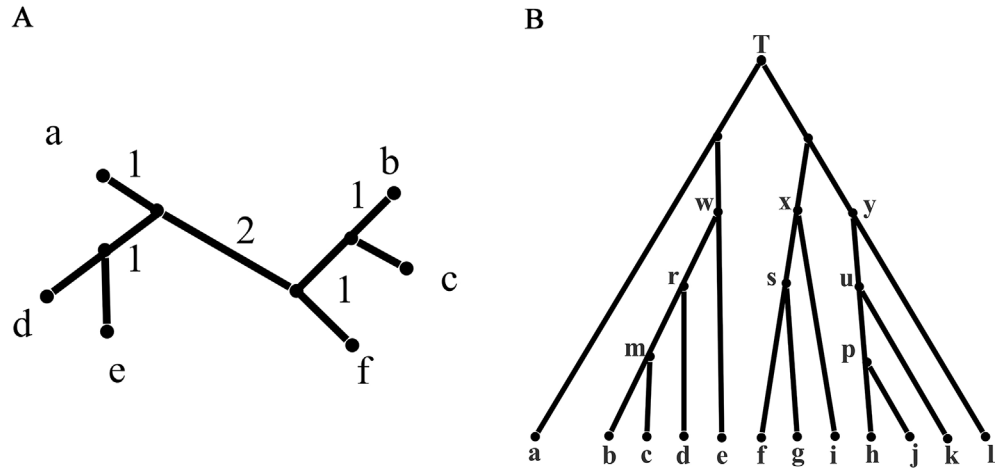


FIGURE 1 | Phylogenetic tree: a phylogenetic tree for illustration and a phylogenetic tree with 12 leaves. **(A)** A phylogenetic tree for illustration with the branch length indicating evolutionary distance. **(B)** A phylogenetic tree with 12 leaves used to generate the first simulation data.

Ever-Expanding Process Based on Quartet and Sextet Weights

As represented by equation W(a|d) = W(a|de) + W(ae|d), there is such an equation W(abc|def) = W(abc|defg) + W(abcg|def), which can be seen as adding a new taxon g to either side of a split abc|def. If W(abc|def) = 0, then W(abc|defg) = 0 and W(abcg|def) = 0. If taxa group $A_1 \subseteq A$ and $B_1 \subseteq B$, or $A_1 \subseteq B$ and $B_1 \subseteq A$, we call the split A|B displays $A_1|B_1$. It is proven in Bandelt and Dress (1992) that $W(A|B) \le W(A_1|B_1)$. Therefore, a split with zero weight cannot be further expanded to larger splits with positive weights.

For a taxa set S with size n, there are $10\binom{n}{6}$ sextets. We first calculate the weights of all quartets and sextets from the MSA, and then we expand them to get all full split weights using an ever-expanding process. Suppose there is a septet of abc|defg type, we have W(abc|defg) = W(abc|def) − W(abcg|def), and there is a similar equation for W(abcg|def), so the weight of W(abc|defg) can be obtained by similar continuous calculations, as follows.

$$\begin{cases} W(abc|defg) = W(abc|def) - W(abcdg|def) \\ W(abcg|def) = W(abg|def) - W(abg|cdef) \\ W(abg|cdef) = W(abg|cde) - W(abfg|cde) \\ W(abfg|cde) = W(afg|cde) - W(afg|bcde) \\ W(afg|bcde) = W(afg|bcd) - W(aefg|bcd) \\ W(aefg|bcd) = W(efg|bcd) - W(efg|abcd) \\ W(efg|abcd) = W(efg|abc) - W(defg|abc) \end{cases}$$

Combining the above equations, we have

$$\begin{aligned} W(abc\,|\,defg) = \tfrac{1}{2}\{&W(abc\,|\,def) - W(abg\,|\,def) + \\ &W(abg\,|\,cde) - W(afg\,|\,cde) + W(afg\,|\,bcd) - \\ &W(efg\,|\,bcd) + W(efg\,|\,abc)\} \end{aligned} \quad (1)$$

For |B| ≥ 4, taking minimum over all possible cases, we have

$$\begin{aligned} W(abc|B) = \max\Big\{&\tfrac{1}{2}\min_{defg\in B}\{W(abc|def) - W(abg|def) + \\ &W(abg|cde) - W(afg|cde) + \\ &W(afg|bcd) - W(efg|bcd) + W(efg|abc)\},0\Big\} \end{aligned} \quad (2)$$

When |A|=4 and |B|=4, the weight of the 4|4 split

$$\begin{aligned} W(A|B) = \min\Big\{&\min_{a\in A}\{W(A-a|B) - W(A-a|B+a)\}, \\ &\min_{a\in B}\{W(A|B-e) - W(A+e|B-e)\} \end{aligned} \quad (3)$$

where A−A′ for two sets A and A′ denotes set difference (subtraction).

For example A={a, b, c, d}, B={e, f, g, h}, there are eight equations for W(abcd|efgh),

$$W(abcd|efgh) = \begin{cases} W(abc|efgh) - W(abc|defgh) \\ W(abd|efgh) - W(abd|cefgh) \\ W(acd|efgh) - W(acd|befgh) \\ W(bcd|efgh) - W(bcd|aefgh) \\ W(abcd|efg) - W(abcdh|efg) \\ W(abcd|efh) - W(abcdg|efh) \\ W(abcd|egh) - W(abcdf|egh) \\ W(abcd|fgh) - W(abcd|fgh) \end{cases}$$

For any split A|B with |A| ≥ 4 and |B| ≥ 4, we traverse the elements in A and B and take out four taxa for each calculation. Suppose a, b, c, d ∈ A and e, f, g, h ∈ B, and we have

$$W(A|B) = \min_{abcd\,\in\,A;\,efgh\,\in\,B}\{W(abcd\,|\,efgh)\} \quad (4)$$

For any 2|n split of ab|B type with c, d, e ∈ B, we calculate their weight by formula (5) referred in Quartet-Net (Yang et al., 2013),

$$W(ab|B) = \max\left\{\frac{1}{2}\min_{cde \in B}\{W(ab|cd) - W(ae|cd) + W(ae|bc) - W(bc|de) + W(ab|de)\}, 0\right\} \quad (5)$$

Finally, for any trivial split of a|S − a type with b, c ∈ S−a in a taxa set S, we calculate the weight as follows (see also Yang et al., 2013):

$$W(a|S-a) = \min_{bc \in S-a}\left\{W(a|bc) - \sum_{a \in A;\, bc \in B} A|B\right\} \quad (6)$$

Formulas (1) – (6) are used to calculate all full splits by decomposing sextet weights iteratively.

The QS-Net Method

QS-Net takes an MSA as input. Suppose that there are n taxa in the taxa set S, which are arranged in the order of 1, 2, 3, …, n. In the initialization step, all triplet, quartet, and sextet weights are calculated directly from the MSAs. We calculate the weights of full splits in the following ways.

- Full split of type A|S − A with |A| ≥ 3 and |S−A| ≥ 3: for the first six taxa—1, 2, 3, 4, 5, and 6—there are 10 sextets. We store these sextets together with their weights in a set X_1. QS-Net then iteratively adds i=7, 8…, n to the left and right parts of the splits stored in X_1 and use equations (2)–(4) to calculate the weights of newly generated splits. Noticing that the only splits that cannot be generated in this way are of type i j k|S_1 − {i, j, k} with j = i − 1, i − 2, …, 2 and k = j − 1, j − 2, …, 1, we calculate their weights using equation (2) and add them to X_1. At the end of each iteration, the splits with a weight of zero are removed because they cannot be further expanded to have a positive weight. After the last iteration, all full splits of type |A| ≥ 3 and |S − A| ≥ 3 have been calculated.
- 2|n – 2 full splits: These splits can be calculated using equation (5). In practice, we use Quartet-Net to calculate their split weights.
- Trivial (1|n − 1) full splits: These splits can be calculated by equation (6).

By the above procedures, we calculate the weights of all full splits. Similar to Yang et al. (2013), it is usually advisable to filter the non-trivial full splits with very low split weights, which tend to be false positives. In practice, we remove splits with weight less than c% of the average weight, where c is a user-defined threshold setting to be 1 in this study. The output file containing all non-zero full splits and their weights is stored in.NEXU file format, which can be visualized using SplitsTree4 (Huson and Bryant, 2006). The time complexity of QS-Net is $O(n^{10})$.

RESULTS AND DISCUSSIONS

To demonstrate QS-Net, we analyzed three artificial data sets and two real data sets. The artificial data sets were generated from a simple tree phylogeny, a phylogenetic scenario with three reticulate events, and a more complicated phylogenetic scenario with five reticulate events. The purpose is to show that the QS-Net method can accurately reconstruct all kinds of evolutionary histories from simple to complicated ones. The real data include a bacterial taxonomy data consisting of 36 bacterial species and the whole genome sequences of 22 H7N9 influenza A viruses downloaded from NCBI influenza database.

The software Dawg (Cartwright, 2005) with model GTR + Gamma + I was used to generate three artificial data sets. The substitution rate is 0.01; the sequence length of the tree is 10,000 bp; the sequence length of the network containing three evolutionary events is 80,000 bp, while the sequence length of the network containing five evolutionary events is 320,000 bp because they are a concatenation of eight and 32 feasible trees. To avoid randomness, we performed 100 Dawg runs on each of the three artificial data sets and applied the 100 MSAs of each data set to QS-Net together with other four popular methods: Quartet-Net (Yang et al., 2013), Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), and Neighbor-Joining (Saitou and Nei, 1987).

Analysis on the Tree Data

The tree data were generated from **Figure 1B** with 12 leaves. For brevity, we only listed reconstructed taxa set in the left or right block containing fewer number of taxa (**Supplementary Material: Table S1**). For example, split bd|acefghijkl was listed as bd. We then normalized each split by the weight of a split successfully constructed by all methods. All trivial full splits were not listed because they can be successfully reconstructed by all five methods. As shown in **Table 1**, all five methods can successfully reconstruct all full splits in the 100 runs of the tree data; the accuracy is equal to the experimental bootstrap value divided by the real bootstrap value. The true-positive split result represents all splits in the real phylogenetic history of the simulated data sets. We listed the number of true-positive splits obtained by the five methods on all simulated data sets in **Table 2**. If a method can reconstruct the true-positive split once in 100 runs, we determined that the true-positive split can be obtained by this method. In addition to true-positive results, other split results reconstructed by the method are false-positive splits, which typically have very few weight values. Except for Neighbor-Joining, the other four methods reconstructed some false-positive splits (here we only list false-positive splits with a bootstrap value ≥10). For example, Quartet-Net and QS-Net reconstruct two additional split al and ae with bootstrap values of 10 and 26, respectively (see **Table 3**). This is because QS-Net and Quartet-Net methods use the same calculation formula for split of 2|n type. Neighbor-Net identifies 35 false-positive splits with bootstrap value ranging from 10 to 40. These false-positive splits may be caused by some random mutations in the tree data set.

TABLE 1 | Comparison of accuracy (the total bootstrap value obtained from the experimental results is divided by the bootstrap BV value) between QS-Net and four other methods.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	100%	100%	100%	100%	100%
Network (3)	100%	100%	70.16%	67.24%	36%
Network (5)	100%	94.74%	58.89%	46.76%	23.68%

Network (3) is the phylogenetic network with three reticulate events, while Network (5) is the phylogenetic network with five reticulate events.

TABLE 2 | The number of true-positive results can be obtained by five methods.

Data set	True	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	9	9	9	9	9	9
Network (3)	25	25	25	21	23	9
Network (5)	38	38	36	30	22	11

The "True" column represents the real number of true-positive splits of the simulated data.

TABLE 3 | The number of false-positive results obtained by five methods.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
Tree	2	2	35	4	0
Network (3)	4	4	16	1	0
Network (5)	4	4	4	1	0

Analysis on the Network Data with Three Reticulate Events

The network data were generated from **Figure 2A** containing three reticulate events A, B, and C, which can be decomposed into eight feasible underlying trees. A feasible tree can be obtained by cutting off one branch respectively at A, B, and C. For example, we can get an underlying tree by cutting off the three edges qA, mB, and oC in the three reticulate events. The sequence data of a taxon m were generated by concatenating partial sequence data from q and partial sequence data from r. All true splits and splits reconstructed by the five methods are listed in **Supplementary Material: Table S2**. The weight of the true split is the sum of the split weights in eight feasible trees. Similarly, we normalized each split with the weight of split ab and multiplied it by 4. As can be seen from the **Table 1**, QS-Net and Quartet-Net accurately reconstructed all true splits in all 100 runs, while Neighbor-Net, Split-Decomposition, and Neighbor-Joining failed to reconstruct a large number of true splits. For example, Neighbor-Net failed to reconstruct split gh, fgi, and fgh in more than 90 runs, and Split-Decomposition was unable to reconstruct split bce and bcde in all 100 runs (**Supplementary Material: Table S2**). Neighbor-Joining obtained even worse result with 16 true splits missing, which is reasonable because Neighbor-Joining only reconstructs trees and retains the strongest compatible splits.

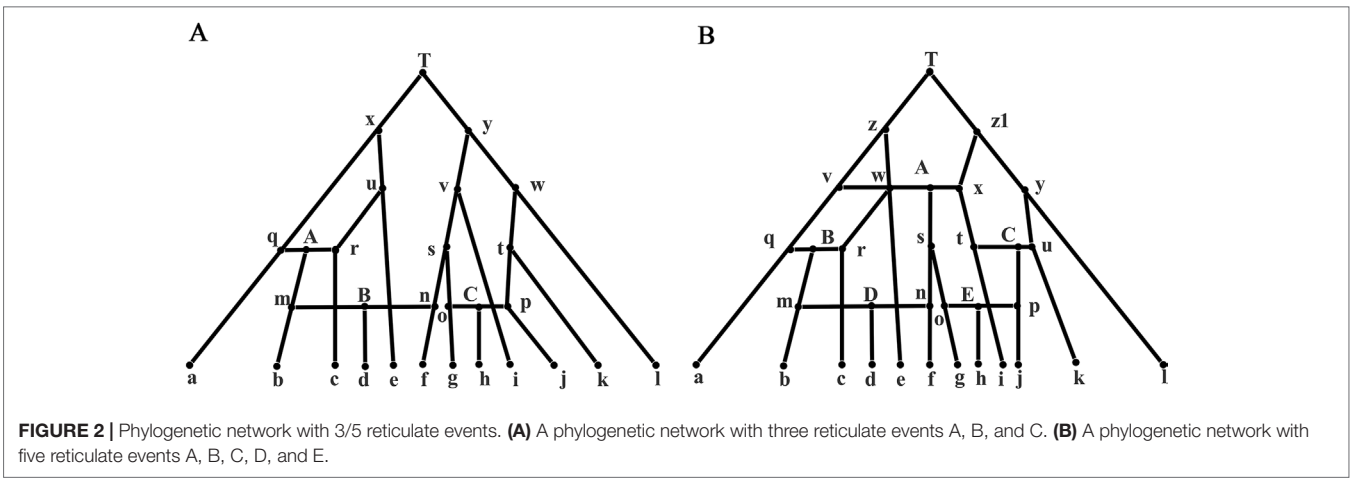


FIGURE 2 | Phylogenetic network with 3/5 reticulate events. **(A)** A phylogenetic network with three reticulate events A, B, and C. **(B)** A phylogenetic network with five reticulate events A, B, C, D, and E.

Analysis on the Network Data with Five Reticulate Events

Supplementary Material Table S3 lists all true splits and splits reconstructed from the five methods from the network data. The data set was generated from **Figure 2B** with a complicated phylogenetic scenario containing five reticulate events. Similarly, the weight of the true split is the sum of the weights of the splits in 32 feasible trees. We normalized each split with the weight of split ce. As can be seen from the **Table 1**, only QS-Net method obtains 100% accuracy in all 100 runs, while the other four methods fail to reconstruct some splits in most runs. For example, Quartet-Net failed in reconstructing split fgi and afg in all 100 runs. In addition to the two splits, Neighbor-Net also cannot reconstruct split hj, bcd, and bcde in more than 90 runs (**Supplementary Material: Table S3**), which happens because Neighbor-Net reduces splits to make the split system planar. Split-Decomposition and Neighbor-Joining still performed poorly. In addition, all methods except for Neighbor-Joining reconstructed some false-positive splits.

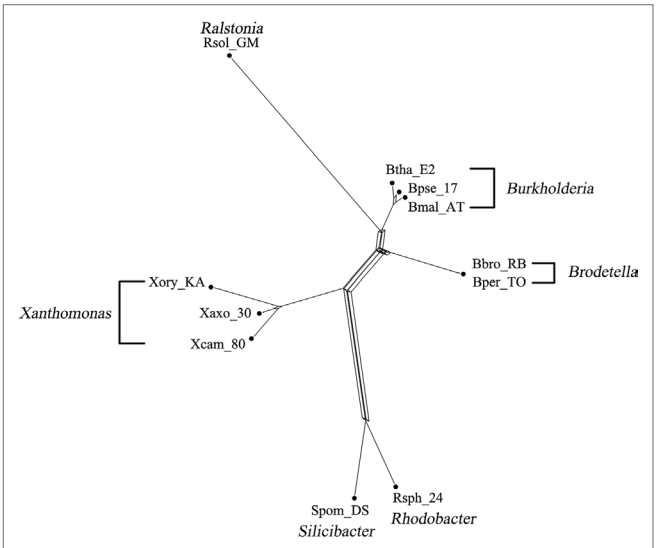


FIGURE 3 | The reconstructed QS-Net network of 11 GC-rich bacteria.

Analysis on the Bacterial Data

The bacterial data set was used in Takahashi and Kryukov (2009) for the analysis of phylogenetic relationships among bacterial species. This data set consists of 36 bacterial genomes containing concatenated sequence of seven genes (16S rRNA, 23S rRNA, gyrB, pyrH, recA, rpoA, and rpoD). The 36 species were divided into three different groups based on different GC content (32–38%, 50–53%, and 64–69%), containing 14, 11, and 11 species, respectively. We took the GC-rich data consisting of 11 bacterial species and a data of 25 species containing both GC-poor and GC-rich bacteria. The MSAs of both data were generated by ClustalW (Larkin et al., 2007) and further fed into to QS-Net, Quartet-Net (Yang et al., 2013), Neighbor-Net (Bryant and Moulton, 2004), Split-Decomposition (Bandelt and Dress, 1992), and Neighbor-Joining (Saitou and Nei, 1987). We ran the program on an MSI laptop with 2.8-GHz processor and 8-GB memory. A comparison of runtime between QS-Net and Quartet-Net on all data sets is shown in **Table 4**; the time statistics for three artificial data sets are the average of all 100 runtimes. The Neighbor-Joining method has the least runtime, and all other three methods can produce results in less than 2 s on all data sets. The reconstructed results were then viewed by SplitsTree4 (Huson and Bryant, 2006). Only three split networks reconstructed by QS-Net and Quartet-Net method on bacterial data set are shown in **Figures 3** and **4**.

Figure 3 shows the phylogenetic network of 11 GC-rich bacterial sequence data set by using QS-Net, which is basically consistent with the experimental results in Takahashi and Kryukov (2009). The reconstructed networks of 25 GC-poor or GC-rich (32–38% and 64–69%) sequence data set reconstructed by QS-Net and Quartet-Net are shown in **Figures 4A, B**, respectively. As can be seen from the figures, the differences between QS-Net and Quartet-Net are quite obvious. There are two distinct parallelograms that represent the reticulate evolution event in the reconstructed network in **Figure 4A** but not in **Figure 4B**, which might be neglected by Quartet-Net due to its inability to identify complicated reticulate events. The numbers of full splits reconstructed by the five methods on bacterial data set and the influenza data set are also listed in **Table 5**. QS-Net constructs a moderate total number of splits among all comparison methods, probably because the full resolution of taxa is not achieved. In the GC-rich data set, Neighbor-Net constructs three more splits than does QS-Net, while in the GC-poor and GC-rich data set, Neighbor-Net constructs 29 more splits than does QS-Net. In addition, by comparing **Figures 3** and **4A**, it can be found that GC content may have an important influence on the evolutionary history of bacteria.

Analysis on the Influenza Data

The data set consisted of the full genome sequence of 22 H7N9 influenza A viruses aligned by ClustalW (Larkin et al., 2007). These viruses have major relations with the H7N9 virus (Gao et al., 2013) that appeared in China in 2013, which caused human mortality. We estimated the phylogenetic relationships

TABLE 4 | A comparison of runtime between QS-Net and Quartet-Net on all data sets.

Method	Tree	Network (3)	Network (5)	GC rich	GC poor and rich	Influenza
QS-Net	1.25 s	6.02 s	24.39 s	0.92 s	9.49 min	3.22 min
Quartet-Net	0.20 s	1.05 s	4.05 s	0.19 s	10.17 s	4.54 s

of these 22 influenza A viruses using Quartet-Net and QS-Net. The results are shown in **Figures 5A**, **B**, respectively. **Table 5** lists the numbers of full splits reconstructed by the five methods on bacterial data set and the influenza data set. General split networks do not actually represent explicit evolutionary events, which makes the interpretation and comparison of reconstruction methods on real data set difficult. So we list the number of splits built by various methods. As can be seen in **Table 4**, QS-Net reconstructs 47 full splits, while Quartet-Net reconstructs 45 full splits.

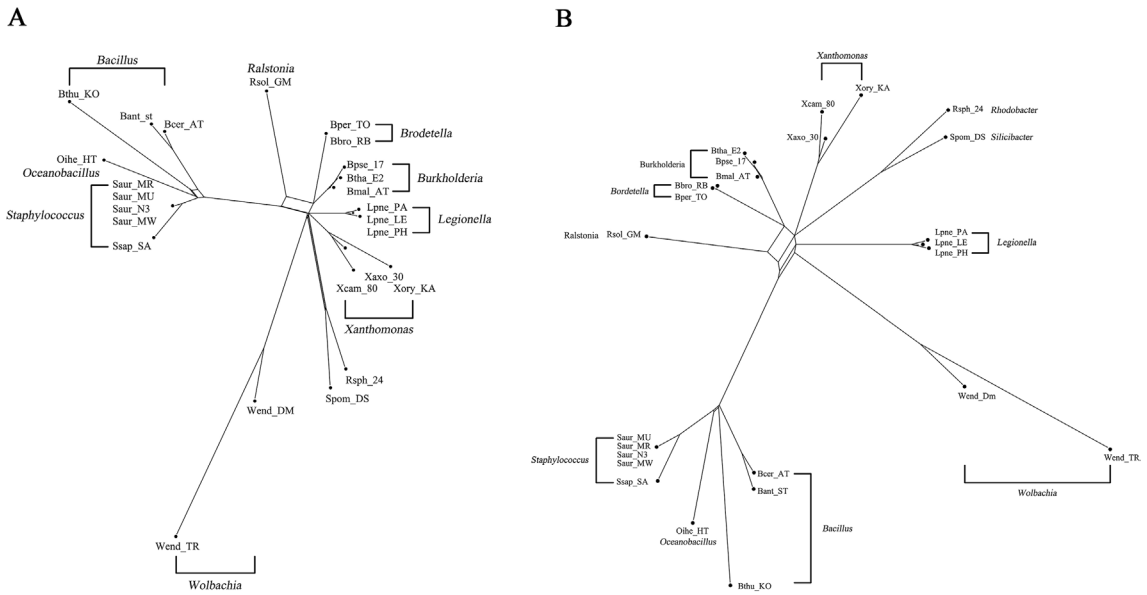


FIGURE 4 | The reconstructed network on 25 GC-poor or GC-rich bacteria. **(A)** The reconstructed QS-Net network of 25 GC-poor or GC-rich bacteria. **(B)** The reconstructed Quartet-Net network of the 25 bacteria.

TABLE 5 | The number of full splits reconstructed by five methods on bacterial data set and the influenza data set.

Data set	QS-Net	Quartet-Net	Neighbor-Net	Split-Decomposition	Neighbor-Joining
GC rich	26	22	29	23	19
GC poor and rich	48	45	77	48	47
Influenza	47	45	68	36	41

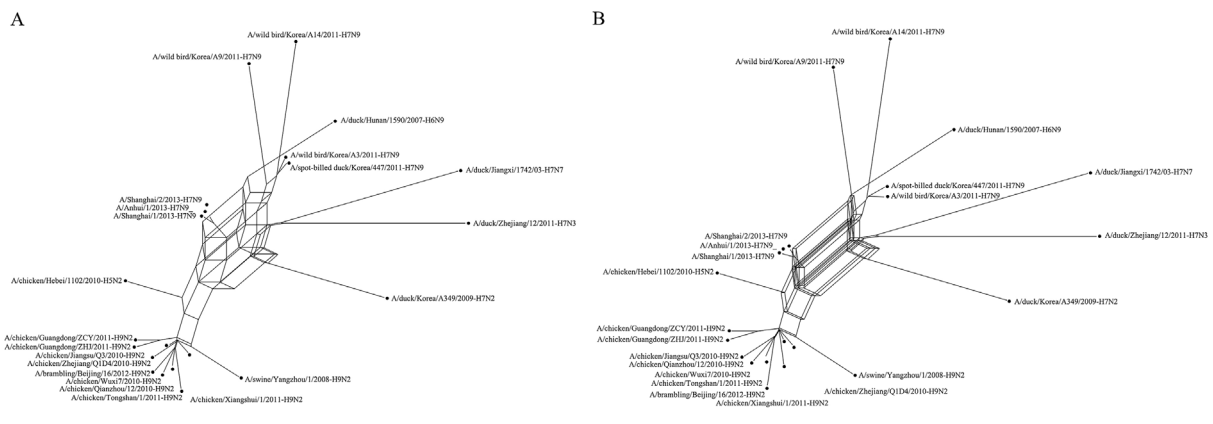


FIGURE 5 | The reconstructed network on influenza data. **(A)** The reconstructed Quartet-Net network related to H7N9 influenza A viruses. **(B)** The reconstructed QS-Net network related to H7N9 influenza A viruses.

The three viruses that caused human death (A/Shanghai/1/2013, A/Shanghai/2/2013, and A/Anhui/1/2013) were combined. The phylogenetic network indicates that these H7N9 viruses may be derived from the reassortment from influenza subtypes, including avian-origin H7N9 viruses, H9N2 viruses, and H7N3 viruses. In **Figure 5B** (constructed by QS-Net), the internal region surrounded by H7N9, H7N7, and H7N3 is more complex than **Figure 5A** (constructed by Quartet-Net), which indicates that the true evolutionary history of H7N9 influenza A viruses is very complex. Of course, the real evolutionary history is unknown, but at least the results constructed by QS-Net are consistent with a few previous findings.

CONCLUSIONS

QS-Net is a method generalizing Quartet-Net. Both simulation studies and real data analyses show that QS-Net has the potential to reconstruct more accurate phylogenetic relationships than its competitors like Quartet-Net and Neighbor-Net. However, the method runs slower than other algorithms, and the major computational difficulty lies in the calculation of 3|4 splits. Nevertheless, the difficulty will be partially resolved with the development of high-speed computers and parallel algorithms. Thus, we believe QS-Net will be useful in identifying more complex reticulate events that will be ignored by other network reconstruction algorithms.

AUTHOR CONTRIBUTIONS

JY and BL conceived the concept of the work and designed the experiments. MT, HL, ZC, and JZ performed literature search. MT, HL, DY, and GT collected and analyzed the data. MT and JY wrote the paper. All authors have approved the final manuscript.

FUNDING

This work was supported by Hainan Provincial Innovation research team (No. 2019CXTD405), National Natural Science Foundation of China (No. 61762034), Hainan Provincial Natural Science Foundation of China (No.618MS057, No.617122) , Hainan Provincial major scientific and technological plans (No. ZDKJ2017012), Natural Science Foundation of Hunan, China (Nos. 2018JJ2461 and 2018JJ3568), New Century Excellent Talents in university (No. NCET-10-0365), National Nature Science Foundation of China (Nos 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, 61672214, and 61702054).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00607/full#supplementary-material

REFERENCES

Avni, E., Cohen, R., and Snir, S. (2015). Weighted quartets phylogenetics. *Syst. Biol.* 64 (2), 233–242. doi: 10.1093/sysbio/syu087

Bandelt, H. J., and Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* 1 (3), 242–252. doi: 10.1016/1055-7903(92)90021-8

Bastide, P., Solís-Lemus, C., Ané, C., William Sparks, K., and Kriebel, R. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst. Biol.* 67 (5), 800–820. doi: 10.1093/sysbio/syy033

Berry, V., and Gascuel, O. (2000). Inferring evolutionary trees with strong combinatorial evidence. *Theor. Comput. Sci.* 240 (2), 271–298. doi: 10.1016/S0304-3975(99)00235-2

Bryant, D., and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21 (2), 255–265. doi: 10.1093/molbev/msh018

Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* 21 Suppl 3 (Suppl_3), iii31. doi: 10.1093/bioinformatics/bti1200

Claudia, S. L., and Cécile, A. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12 (3), e1005896. doi: 10.1371/journal.pgen.1005896

Eslahchi, C., Habibi, M., Hassanzadeh, R., and Mottaghi, E. (2010). MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method. *BMC Evol. Biol.* 10 (1), 254. doi: 10.1186/1471-2148-10-254

Fenderson and Bruce, (2008). Molecular biology of the cell, 5th Edition. *Shock* 30 (1), 100. doi: 10.1097/01.shk.0000286288.33338.f6

Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., et al. (2013). Human infection with a novel avian-origin influenza A (H7N9) virus. *N. Engl. J. Med.* 368 (20), 1888. doi: 10.1056/NEJMoa1304459

Grünewald, S., Forslund, K., Dress, A., and Moulton, V. (2007). QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol. Biol. Evol.* 24 (2), 532–538. doi: 10.1093/molbev/msl180

Grunewald, S., Spillner, A., Bastkowski, S., Bogershausen, A., and Moulton, V. (2013). SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10 (1), 151–160. doi: 10.1109/TCBB.2013.8

Huber, K. T., Langton, M., Penny, D., Moulton, V., and Hendy, M. (2002). Spectronet: a package for computing spectra and median networks. *Appl. Bioinformatics* 1 (3), 159–161. doi: 10.1385/ABAB:97:1:01

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* 44 (1), 17–48. doi: 10.1093/sysbio/44.1.17

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23 (2), 254–267. doi: 10.1093/molbev/msj030

Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511974076

Keijsper, J. C. M., and Pendavingh, R. A. (2014). Reconstructing a phylogenetic level-1 network from quartets . *Bull. Math. Biol.* 76 (10), 2517–2541. doi: 10.1007/s11538-014-0022-z

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23 (21), 2947–2948. doi: 10.1093/bioinformatics/btm404

Leducq, J.-B., Henault, M., Charron, G., Nielly-Thibault, L., Terrat, Y., Fiumera, H. L., et al. (2017). Mitochondrial recombination and introgression during speciation by hybridization. *Mol. Biol. Evol.* 34 (8), 1947–1959. doi: 10.1093/molbev/msx139

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32 (10), 2798–2800. doi: 10.1093/molbev/msv150

Nelson, M. I., Viboud, C., Simonsen, L., Bennett, R. T., Griesemer, S. B., George, K. S., et al. (2008). Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog.* 4 (2), e1000012. doi: 10.1371/journal.ppat.1000012

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405 (6784), 299–304. doi: 10.1038/35012500

Park, H. J., Jin, G., and Nakhleh, L. (2010). Bootstrap-based support of HGT inferred by maximum parsimony. *BMC Evol. Biol.* 10 (1), 131. doi: 10.1186/1471-2148-10-131

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8 (11), e1002967. doi: 10.1371/journal.pgen.1002967

Ranwez, V., and Gascuel, O. (2002). Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Mol. Biol. Evol.* 19 (11), 1952–1963. doi: 10.1093/oxfordjournals.molbev.a004019

Rieseberg, L. H., Baird, S. J. E., and Gardner, K. A. (2000). Hybridization, introgression, and linkage evolution. *Plant Mol. Biol.* 42 (1), 205–224. doi: 10.1023/A:1006340407546

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406. doi: 10.1093/oxfordjournals.molbev.a040454

Sayyari, E., and Mirarab, S. (2016). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics* 17 (Suppl 10), 101–113. doi: 10.1186/s12864-016-3098-z

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34 (12), 3292–3298. doi: 10.1093/molbev/msx235

Takahashi, M., and Kryukov, K. N. (2009). Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 93 (6), 525–533. doi: 10.1016/j.ygeno.2009.01.009

Tria, F., Caglioti, E., Loreto, V., and Pagnani, A. (2010). A stochastic local search algorithm for distance-based phylogeny reconstruction. *Mol. Biol. Evol.* 27 (11), 2587. doi: 10.1093/molbev/msq154

Vachaspati, P., and Warnow, T. (2015). ASTRID: accurate species trees from internode distances. *BMC Genomics* 16 (10), S3. doi: 10.1186/1471-2164-16-S10-S3

Van Iersel, L., Kelk, S., Rupp, R., and Huson, D. (2010). Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26 (12), i124–i131. doi: 10.1093/bioinformatics/btq202

Vijaykrishna, D., Mukerji, R., and Smith, G. J. D. (2015). RNA virus reassortment: an evolutionary mechanism for host jumps and immune evasion. *PLoS Pathog.* 11 (7), e1004902. doi: 10.1371/journal.ppat.1004902

Wen, D., Zhu, J., Yu, Y., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* 67 (4), 735–740. doi: 10.1093/sysbio/syy015

Wheeler, T. J. (2009). *Large-scale Neighbor-Joining with NINJA*. Berlin Heidelberg: Springer. doi: 10.1007/978-3-642-04241-6_31

Willson, S. J. (2013). Reconstruction of certain phylogenetic networks from their tree-average distances. *Bull. Math. Biol.* 75 (10), 1840–1878. doi: 10.1007/s11538-013-9872-z

Yang, J., Grünewald, S., and Wan, X. F. (2013). Quartet-Net: a quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30 (5), 1206–1217. doi: 10.1093/molbev/mst040

Yu, Y., and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16 (10), S10. doi: 10.1186/1471-2164-16-S10-S10

Yu, Y., Nakhleh, L., Than, C., and Degnan, J. H. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60 (2), 138–149. doi: 10.1093/sysbio/syq084

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2017). Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* 35 (2), 504–517. doi: 10.1101/124982

Conflict of Interest Statement: Authors DY and GT were employed by company Geneis (Beijing) Co. Ltd. The remaining authors declare no competing interests.

Copyright © 2019 Tan, Long, Liao, Cao, Yuan, Tian, Zhuang and Yang. 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) and the copyright owner(s) 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 Physiology

ORIGINAL RESEARCH
published: 24 September 2019
doi: 10.3389/fphys.2019.01160

A Computational Pipeline for the Extraction of Actionable Biological Information From NGS-Phage Display Experiments

Antonios Vekris[1†], Eleftherios Pilalis[2,3†], Aristotelis Chatziioannou[2,3] and Klaus G. Petry[1]*

[1] UMR 1049 and U1029, INSERM, Bordeaux, France, [2] Metabolic Engineering and Bioinformatics Program, Institute of Chemical Biology, National Hellenic Research Foundation, Athens, Greece, [3] eNIOS Applications P.C., Athens, Greece

Phage Display is a powerful method for the identification of peptide binding to targets of variable complexities and tissues, from unique molecules to the internal surfaces of vessels of living organisms. Particularly for *in vivo* screenings, the resulting repertoires can be very complex and difficult to study with traditional approaches. Next Generation Sequencing (NGS) opened the possibility to acquire high resolution overviews of such repertoires and thus facilitates the identification of binders of interest. Additionally, the ever-increasing amount of available genome/proteome information became satisfactory regarding the identification of putative mimicked proteins, due to the large scale on which partial sequence homology is assessed. However, the subsequent production of massive data stresses the need for high-performance computational approaches in order to perform standardized and insightful molecular network analysis. Systems-level analysis is essential for efficient resolution of the underlying molecular complexity and the extraction of actionable interpretation, in terms of systemic biological processes and pathways that are systematically perturbed. In this work we introduce PepSimili, an integrated workflow tool, which performs mapping of massive peptide repertoires on whole proteomes and delivers a streamlined, systems-level biological interpretation. The tool employs modules for modeling and filtering of background noise due to random mappings and amplifies the biologically meaningful signal through coupling with BioInfoMiner, a systems interpretation tool that employs graph-theoretic methods for prioritization of systemic processes and corresponding driver genes. The current implementation exploits the Galaxy environment and is available online. A case study using public data is presented, with and without a control selection.

Keywords: phage display, Galaxy platform, enrichment analysis, network analysis, biological interpretation, Reactome, Gene Ontology

OPEN ACCESS

Edited by:
Marco Antoniotti,
University of Milano-Bicocca, Italy

Reviewed by:
Marco Vanoni,
University of Milano-Bicocca, Italy
Lydie Lane,
Swiss Institute of Bioinformatics (SIB), Switzerland

***Correspondence:**
Aristotelis Chatziioannou
achatzi@eie.gr

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *23 November 2018*
Accepted: *28 August 2019*
Published: *24 September 2019*

Citation:
Vekris A, Pilalis E, Chatziioannou A and Petry KG (2019) A Computational Pipeline for the Extraction of Actionable Biological Information From NGS-Phage Display Experiments. Front. Physiol. 10:1160. doi: 10.3389/fphys.2019.01160

INTRODUCTION

Phage Display has been widely used to select peptides binding to a variety of targets, *in vitro* or *in vivo*, with complexities ranging from a single macromolecule (Rodi and Makowski, 1999; Bábíčková et al., 2013) to diffuse pathological lesions (Pasqualini and Ruoslahti, 1996). Peptides identified using this technique have been successfully used for specific site drug delivery and *in vivo*

imaging (Deutscher, 2010). The complexity of the selected repertoires of peptides is a function of the complexity of the target. Complex selections were poorly analyzed before the introduction of Next Generation Sequencing (NGS), which offered a detailed view of the peptide sequences (Dias-Neto et al., 2009). Software solutions were developed to compare the contents of several repertoires to identify common or specific sequences (Kolonin et al., 2006). In parallel, the hypothesis of mimicked proteins was advanced, based on the assumption that some peptides share sequence similarity with protein domains, and thus mimic the physiological interaction of the protein domain with its targets. In this scope, sequence comparison was usually performed using available tools, performing probabilistic (BLAST) (Altschul et al., 1990) or best-match (Needleman–Wunsch) (Smith and Waterman, 1981) mappings. A tool more adapted to analysis of phage display data, named PepTeam was developed by Hume et al. (2013), based on an algorithm producing all the ungapped matches of the peptides of a repertoire, compared against a set of proteins. Here we introduce PepSimili, a new computational tool significantly extending the capabilities of PepTeam and suitable for large-scale analysis of phage display data derived from NGS. PepSimili integrates the mapping function of PepTeam and extends the analysis with (a) an evaluation and subtraction of the local noise due to random mappings, (b) the subtraction of the signals produced by a control repertoire, and (c) filtering of derived proteins using a mapping score.

Moreover, PepSimili automatically manages a systems-level biological interpretation, in terms of underlying biological processes and master regulator genes. Pathway and functional analysis is performed by coupling the mapping functions of PepSimili with BioInfoMiner (Pilalis and Chatziioannou, 2013; Koutsandreas et al., 2016), an algorithm that performs systemic functional interpretation of the phenotype interrogated through the phage-display experiment. The interpretation algorithm performs by projecting the highest ranked proteins onto ontological and pathway networks with hierarchical structure. Highest ranked proteins are those presenting statistically significant accumulation of non-random peptide sequence matches. Their mapping on ontological and pathway networks enables the extraction of ranked lists of putative systemic processes and/or pathways, reflecting the underlying components involved in the manifestation of the investigated phenotype. The master regulatory genes and their respective protein products are ranked according to their contribution to the systemic processes.

Overall, PepSimili derives a systems-level interpretation of the mechanisms impacted by the cumulative effect of multiple mimicking peptides on protein networks. Ultimately, it manages to shortlist and rank candidate target proteins deriving from Phage Display experiments, according to their functional impact. The application is implemented on an instance of the Galaxy platform (Afgan et al., 2018). Through its user-friendly visual editor, the execution of the workflow is easily accessible to the basic user without prior experience in bioinformatics or in command-line oriented analyses. Additionally, the Galaxy platform already provides the tools necessary for the manipulation of the raw fastq files including quality filtering, trimming of the sequences to isolate the variable part of the recombinant phages and DNA to protein translation. PepSimili is the first application for the phage display technology implemented in Galaxy and which provides efficient mapping of short peptides on whole proteome databases. The tool is available online at http://pepsimili.e-nios.com:8080.

MATERIALS AND METHODS

Workflow Implementation

The workflow, outlined in **Figure 1**, is written in Python language and implemented as a tool in a Galaxy cloud platform (Afgan et al., 2018).

Inputs

PepSimili is presented as an integrated tool in Galaxy, accepting as inputs:

- A Test repertoire, containing the sequences of the peptides, with a length ranging from 5 to 15 residues, covering all commercially available phage display libraries (the most common being of 7–12 residues).
- A Control repertoire, if available. If not available, a file with a single poly-Tyr of the same length as the peptides (e.g., "WWWWWWW," for $n = 7$) can be used.
- A table with the distribution of the amino acids (percentages) in the library used for the selections.
- A fasta file of the proteome, or a subset of proteins of interest.
- The threshold of similitude, h.
- The minimal z-score to be considered as significant for the selection of the outliers.
- The *p*-values and corrected *p*-values for BioInfoMiner.

Workflow Steps

The main steps of the workflow are the following:

Calculation of the Amino Acid Frequencies of the Test Repertoire

The respective frequency of each amino acid in the library is calculated as a percentage.

Building of a Mock Repertoire

A Mock repertoire is built, composed of peptides of the same length and number (unique) as the peptides of the Test repertoire. Peptide sequences are quasi random, but respecting the amino acid frequencies of the phage library, as calculated in step 1. The Mock repertoire is used for the estimation of the noise produced by random mappings.

Mapping of Test and Control and Mock Repertoires on the Proteome

The problem of mapping a set of peptides on a set of proteins, respecting a threshold of similarity h, was previously addressed in Hume et al. (2013). Similarity between two peptides is evaluated using the PAM30 substitution matrix (Jones et al., 1992).

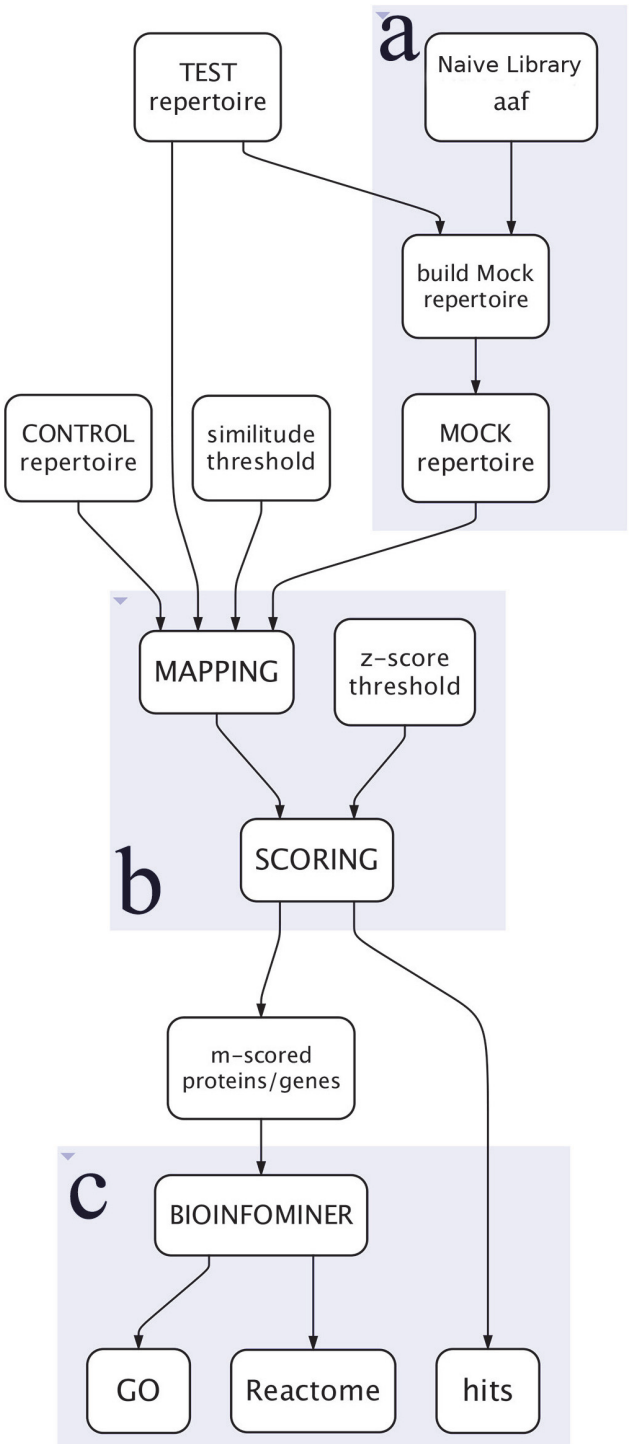


FIGURE 1 | Simplified view of the workflow. **(A)** Represents the construction of the Mock repertoire, **(B)** the mapping and scoring processes, and **(C)** the BioInfoMiner analysis of the proteins/genes of interest.

The algorithms, producing all the ungapped matches, of the peptides of a repertoire, on a set of proteins, were implemented in C++ and the code of the four modules necessary to produce the mappings and provide the resulting profiles is available at https://github.com/cbib/pepteam.

In total three mappings are performed, respectively for Test (T), Control (C) and Mock (M) repertoire (built in step 2). In the resulting files are reported the matching peptides, the similarity score and the matching position on the corresponding protein.

Signal Extraction

The mappings are used to produce a signal profile of mappings for each protein and for each of the T, C, and M repertoires. As signal profile is defined the sum of the hits in each amino acid position on the protein. The profile of background noise, as estimated from the Mock repertoire and representing random mappings, is subtracted from the signal profile of the Test repertoire, for each protein. If a Control repertoire is available, corresponding profiles are subtracted too, in order to extract a final signal profile representing meaningful peptide matches.

Scoring and Ranking of Proteins

After subtraction, the resulting signal profiles are used to generate a mapping score for each protein, termed m-score, which is the sum of the mappings from all positions, divided by the portion of the protein comprising at least one peptide match.

The distribution of the *m*-scores is calculated and each protein is annotated as *z*-score. The *z*-score cut-off set by the user (confidence level) is used to extract the list of proteins of interest for the next step of the analysis, which are thus outliers according to the calculated m-score distribution.

Systemic Biological Interpretation

Biological interpretation is performed for the set of promoted proteins from the previous step, using adapted algorithms from Chatziioannou and Moulos (2011), Moutselos et al. (2011), Pilalis and Chatziioannou (2013). The algorithm performs statistical and network analysis on controlled biological hierarchical vocabularies, here Gene Ontology (GO) (Ashburner et al., 2000) and Reactome pathways (Croft et al., 2014). This step (see section "Graph-Based Biological Interpretation," below) derives significantly impacted biological processes and the respective driver genes linking these processes. It should be noted here that, with minimal programming effort, this algorithmic step can be adapted to exploit additional biological ontologies for network analysis.

Outputs

The workflow produces as outputs:

- The signal profiles of the mapping of the peptides of the Test repertoire, showing the largest cluster of similar peptides, each peptide belongs to (#5).
- The signal profiles of the mappings of the Test, Mock and Control repertoires (#6, #7, and #8, respectively).
- The signal profiles of the mappings of the Test repertoire, minus both the Mock random mappings and the Control mappings (#9).
- Two files corresponding to the Test and Mock repertoires (#10 and #11, respectively) reporting all the mappings on each protein, indicating the position of the mapping,

the homology of the peptide with the underlying protein segment, the peptide sequence, and three characteristics of the peptide: the number of occurrences in the repertoire, the number of peptides in the biggest cluster of similar peptides in the Test repertoire and the number of proteins on which the peptide is mapping.
- The list of proteins of interest, with z-scores above the chosen threshold (#12).
- The list of genes of interest, encoding the proteins of interest (#13), used as input for BioInfoMiner.
- The enriched GO terms characterizing the genes of interest (#14).
- The list of driver genes linking the processes pinpointed by the GO terms enrichment (#15).
- The list of the Reactome pathways, in which the genes of interest participate (#16).
- The list of the Reactome gene ranking (#17).

Parameters

PepSimili is an integrated tool performing all steps of the analysis of the available repertoires in a single run. A group of satellite and intermediate scripts are also available in the platform, for access to intermediate steps of the workflow (listed in the tool menu as "PepSimili tools").

Size of the Experimental Repertoires

The quality of the experimentally obtained peptide repertoires determines the mappings and thus the m-scoring. PepSimili uses the unique peptide sequences present in a repertoire to produce the mappings and calculate the m-scores. It is necessary to dispose of repertoires large enough to optimize the density of the mappings. A minimum of 40000 unique peptide sequences of 7mers is necessary. Usually NGS provides millions of reads for each repertoire, corresponding at hundreds of thousands of unique peptide sequences for targets as simple as cell cultures, thus covering the requirements of PepSimili.

For a visualization of the distribution of the *m*-score the script Sat_distri can be used; it produces a table with the distribution of the scores from any profile file.

Some experimental conditions may produce Test and Control repertoires of different sizes. If the Control repertoire does not fulfill the required equal size, it is advisable to complete by an equal number of unique sequences randomly chosen from the Test repertoire.

Amino Acids Distribution and Mock Repertoire

The distribution along the proteins of the mappings is sensitive to the frequencies of the amino acids of each repertoire (aaf), depending on local distributions. To generate a Mock repertoire allowing an adapted evaluation of the local random noise, it is necessary to apply an amino acid distribution as the one observed in the library being used for the selections. Most of the libraries are constructed using NNK codons, and the distribution of codons and amino acids is further distorted during the amplification of the library. Usually this information is available for commercially available libraries. For custom constructs it is necessary to include a sample of the library in the NGS experiments and calculate the aaf table. The script Sat_aaf uses as input a peptides occurrences table and produces the corresponding amino acids frequencies file. At minimum, if no experimental information is available, it is advised to use the theoretical amino acids frequencies.

The influence of the amino acids frequencies on the mapping's distribution is shown in **Supplementary Figures 1**, **2**. PepSimili accepts as input aaf tables with frequencies expressed as percentages.

Similarity and h Threshold

The degree of similarity of two peptides is calculated using the PAM30 substitution matrix. Only positive values can be handled by PepSimili, ranging from 0 to 1. For 7mers we recommend a threshold h between 0.4 and 0.8, depending on the desirable degree of similarities. As the evaluation of the random mappings is made using the same stringency and this background noise is systematically subtracted from the Test repertoire signals, high stringencies are not obligatory. On the contrary, when the Test repertoire contains a low number of unique sequences, it is advisable to decrease the threshold h, in order to accordingly increase the density of the mappings and obtain a more suitable m-score distribution, for the selection of the proteins of interest. The influence of the choice of the h threshold on the profiles is shown in **Supplementary Figure 3**.

Confidence Level Threshold (*Z*-Score)

The distribution of the m-scores is the function of the number of peptides in the repertoires and the threshold h chosen for the analysis. When both the random and the control (non-specific ones) profiles are subtracted from the test profiles, theoretically all remaining signal is significant and there is no need to focus on outliers of the distribution of the m-scores (**Supplementary Figure 4**). However, it is advisable to approach the Y = aah; described system by the selection(s) starting with proteins presenting the relatively highest m-scores, and the default *z*-score threshold of 2.58 usually selects sufficient numbers of proteins to build a first overview of the system under study, using BioInfoMiner's outputs. A threshold as low as 1.5 is still significant given the distribution of the *m*-scores.

Limitations

Proteomes and Fasta File Header Format

It is mandatory to use a particular header format for the proteins fasta file, so the script correctly extract the gene symbols for biological interpretation: >ENSGXXXXXXXX| ENSTXXXXXX| Gene Symbol| ENSPXXXXXXXX. Such files can be obtained from BioMart (Smedley et al., 2015), using a simple query with, as filters, Gene stable ID, Transcript stable ID, Gene symbol and Protein stable ID. Two sets of human proteins are already included, one of general use (proteome 20k) and one minimal (proteome 10k), more restricted to interaction molecules, adapted for selections performed *in vivo* or on cells accessible via the blood vessels (e.g., endothelial cells such as HUVEC). These proteins belong in either of the following three classes: plasma membrane, extracellular matrix, or secreted proteins, and were selected based on their annotation with

GO terms Cellular localisation (GO:0051641) and Extracellular Region Part (GO:0044421), in addition to the proteins of the human plasma proteome, taken from the Human Plasma Peptide Atlas (Schwenk et al., 2017). For each protein isoform, the most complete in terms of exons was included.

However, the user can upload and use any proteome respecting the above-mentioned header format and perform manually the biological interpretation, using the BioInfoMiner module available on the PepSimili server. The functional analysis currently supports *Homo sapiens*, *Mus Musculus*, *Rattus norvegicus*, *Gallus gallus*, *Sus scrofa*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*.

Test Repertoire Size

We mentioned the necessity of a Test repertoire with more than 40000 unique sequences in order to generate optimal results from our tool. It is possible to generate partial results for smaller repertoires, which are often obtained for relatively simple target systems (assuming fewer binding sites). A satellite script named Sat_scoring is provided, using as input a profile file, preferably the T-M-C one, and producing a table of *m*-scored proteins.

Graph-Based Biological Interpretation

BioInfoMiner algorithm uses protein annotations and ontologies as a starting point for functional and pathway analysis with statistical and graph-theoretic methods (Chatziioannou and Moulos, 2011; Moutselos et al., 2011; Pilalis and Chatziioannou, 2013; Koutsandreas et al., 2016). The algorithm comprises two main steps:

Ontological Process and Pathway Prioritization

The algorithm employs a combination of a parametric (Hypergeometric) and a non-parametric statistical test (bootstrap resampling). Initially the Hypergeometric test is used to assess the over-representation and initial ranking of annotation terms in the input gene list. This ranking is corrected by performing bootstrapping as an alternative to multiple test correction methods (Bonferroni, FDR), thus avoiding false assumptions about the distribution of *p*-values. Instead of adjusting the *p*-values, the bootstrapping algorithm reorders the initial distribution and prioritizes the less frequently observed enrichments which tend to represent broader pathways or functions and, thus, are of stronger biological content (Pilalis and Chatziioannou, 2013; Pilalis et al., 2015).

Gene/Protein Prioritization

Gene prioritization is performed by a graph-theoretical approach that exploits an ontological direct acyclic graph structure to detect and rank genes according to their impact as linkers in the topology of that graph (Moutselos et al., 2011; Koutsandreas et al., 2016), using semantic measure techniques. As background graphs, are used variations of the following ontologies and hierarchical pathways, corrected for inconsistencies (annotation bias, information content imbalance, gaps): Gene Ontology, Reactome, MGI Mammalian Phenotype Ontology (annotation of Human genes) and Human Phenotype Ontology (HPO).

The extracted ranked list of systemic processes and/or pathways, reflects the underlying components involved in the manifestation of the investigated phenotype, and provides a descriptive snapshot that links and integrates the examined individual genes into broader functional, indispensable modules that shape the cellular phenotype. The master regulatory genes and their respective protein products are ranked according to their contribution to the systemic processes.

RESULTS

Galaxy-Based Tool for Integrated Analysis of Phage Display Data

The Galaxy instance of PepSimili is available at http://pepsimili.e-nios.com:8080. The tool is easily accessible in the left-hand menu ("Tools"), under the section *pepSimili*. There are two additional tool sections, *PepSimili tools* and *PepSimili subworkflows*, which comprise the collection of satellite/intermediary scripts and partial worklows for mapping repertoires and scoring proteins, respectively.

PepSimili main input and outputs are shown in **Figure 2**. The tool accepts a Test and a Control repertoire, the amino-acid frequencies of the phage display library, the homology and confidence cut-offs and the *p*-values for the enrichment analysis (see section "Materials and Methods") (**Figure 2A**). All steps are executed automatically, including the biological interpretation.

In **Figures 2B,C** are shown screenshots of, respectively, the output proteins with their m-scores, and an example of a file reporting peptide hits, including their headers, total hit similarity and the list of peptides matching the protein sequence, including additional details and metrics for each peptide (see Materials and Methods).

In **Figure 2D** is shown a heatmap, depicting the mapping of prioritized genes to systemic processes. The rank of each gene depends on the number of processes to which a gene participates. The more processes a gene is mapped to, the higher its rank is, highlighting its importance as a regulatory hub on the ontological network.

Figure 3 shows the peptide mappings on a segment of a protein (WASF1, see Case Study further below), also illustrating the calculation of the aah-score at each position, which is defined as the sum of the total peptide matches.

Case Study

We present a case study using published data of phage display repertoires (Brinton et al., 2016). We used two of the described repertoires. Both were selected *in vitro*, on HUVEC cells that were cultured either in normal medium (Control) or in tumor conditioned medium (tcm) by tumoral cells (pancreatic adenocarcinoma, PDAC), which is serving as Test. The two samples of cultured cells were separately used for the biopanning with a combinatorial library of cyclic 7mers on phage display. In these studies, the recombinant portion of the pIII of the phage was amplified by PCR and used as template for deep sequencing. The aim of the study was to identify binders specific of the tcm-treated endothelial cells, expecting that their targets would

A

B

1
average m-score: 1.00344
stdev: 0.029515
thr: 1.03296 at 1.0 stdevs
9 proteins retained
ENSG00000278466|ENST00000630886|MUC2|ENSP00000486476 1.06452
ENSG00000072195|ENST00000312358|SPEG|ENSP00000311684 1.03448
ENSG00000204842|ENST00000377617|ATXN2|ENSP00000366843 1.06667
ENSG00000152217|ENST00000282030|SETBP1|ENSP00000282030 1.06667
ENSG00000215182|ENST00000621226|MUC5AC|ENSP00000485659 1.12579
ENSG00000113615|ENST00000398844|SEC24A|ENSP00000381823 1.30435
ENSG00000183484|ENST00000329797|GPR132|ENSP00000328818 1.2
ENSG00000117983|ENST00000529681|MUC5B|ENSP00000436812 1.09272
ENSG00000006377|ENST00000518156|DLX6|ENSP00000428480 1.42857

C

1	2	3
protein_id	similarity	peptide
ENSG00000006377|ENST00000518156|DLX6|ENSP00000428480	293	42.45.QEPPPLPPPPFPQLH.1.1.10,48.40.TPPITPQPHSSQPPP.1.1.3
ENSG00000113615|ENST00000398844|SEC24A|ENSP00000381823	1093	112.41.ASQNPTTTSKLATPF.1.1.1,120.47.PMPAHSFLPPPSLPH.1.1.1
ENSG00000183484|ENST00000329797|GPR132|ENSP00000328818	380	25.41.SLGLSHHAVNHISFE.1.1.1,15.40.SPLSPTAPWTPLRLS.1.1.2
ENSG00000152217|ENST00000282030|SETBP1|ENSP00000282030	1596	291.43.PEPSSHDAPSAPARS.1.1.1,1525.44.QEPPPLPPPPFPQLH.1.1.10,1
ENSG00000215182|ENST00000621226|MUC5AC|ENSP00000485659	5654	4101.43.PQVSTASPQTTITIS.1.1.1,2432.41.PQVSTASPQTTITIS.1.1.1,281
ENSG00000131061|ENST00000375200|ZNF341|ENSP00000364346	854	194.44.PHPTLPPNQALDPPG.1.1.2,183.41.QEPPPLPPPPFPQLH.1.1.10
ENSG00000204842|ENST00000377617|ATXN2|ENSP00000366843	1313	208.42.PAATPTPPSTRTSPS.1.1.11,655.51.PAATPTPPSTRTSPS.1.1.11,182
ENSG00000278466|ENST00000630886|MUC2|ENSP00000486476	2817	1466.42.PAATPTPPSTRTSPS.1.1.11,1473.42.PAATPTPPSTRTSPS.1.1.11,1
ENSG00000117983|ENST00000529681|MUC5B|ENSP00000436812	5762	3085.41.ATPMAPTSTMHPSHM.1.1.1,5269.42.IYSPASPHPSTPSPT.1.1.1,25
ENSG00000072195|ENST00000312358|SPEG|ENSP00000311684	3267	2794.40.HRSPGSPNAMGPPLA.1.1.1,2808.43.PAATPTPPSTRTSPS.1.1.11

D

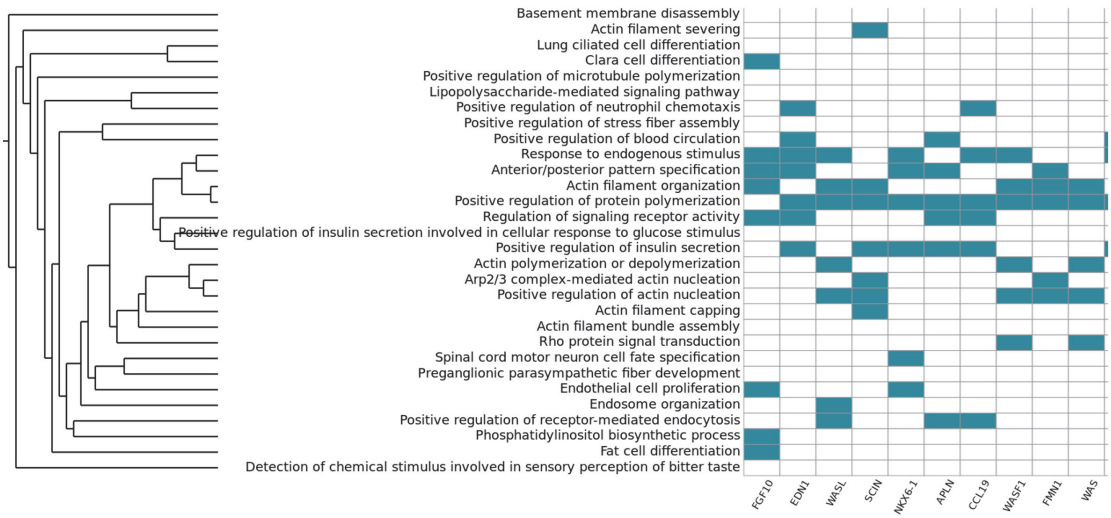


FIGURE 2 | PepSimili main input and outputs. **(A)** The input form of the PepSimili automated workflow. Confidence level refers to the *z*-score cut-off, **(B)** output proteins, ranked by *m*-score, **(C)** output tabular file reporting the hits, **(D)** a heatmap visualizing the mapping of prioritized genes (*x* axis) to systemic processes (*y* axis).

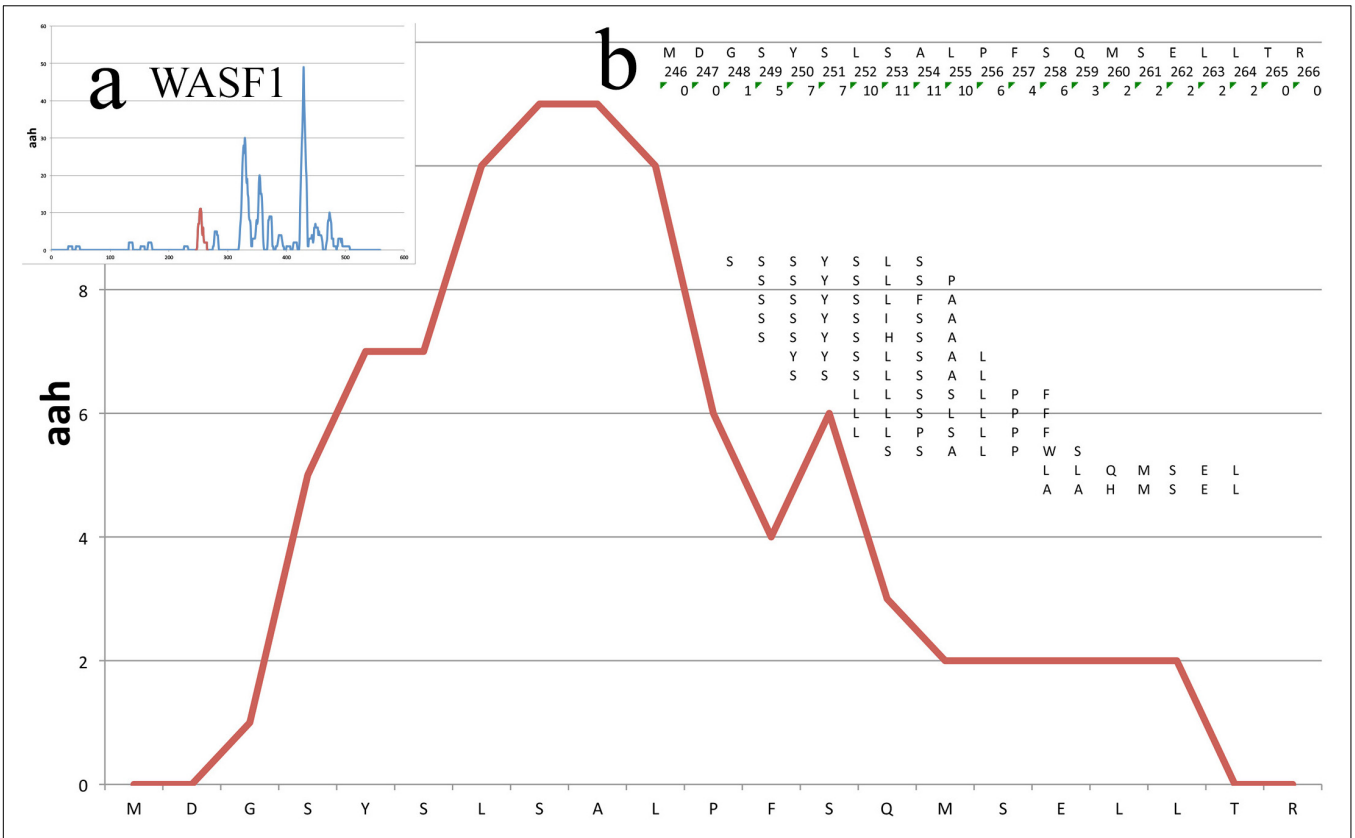


FIGURE 3 | Mappings on a small portion of the WASF1 protein. The profile of the mappings of the HUVEC_tcm peptides (HUVEC under tumor conditioned medium) on the WASF1 protein is shown here, between residues 246 and 266. The *x*-axis presents the sequence, the *y*-axis the amino acid hits. Insert **(a)** shows the position of this particular stretch on the complete profile of the protein (red peek). Insert **(b)** shows in the first row the amino acid sequence, in second row the position of the amino acids and in third row the total aah per residue. Further below are shown the 13 peptides defining this stretch, positioned in frame with their similar sequence on the protein.

be also expressed *in vivo* by cells in the microenvironment of cancerous tumors. Pepsimili extends the scope of the study to the identification of proteins with subsequences similar to the selected peptides. The full run is online available at http://pepsee.e-nios.com:8080/u/avek/h/example-huvectcm-vs-huvec.

Table 1 summarizes the genes prioritized by the BioInfoMiner algorithm using Gene Ontology and Reactome pathways. The gene prioritization results are shown in **Supplementary Table 1** (GO) and **Supplementary Table 2** (Reactome). Interestingly, the highest ranked genes include known markers of PDAC, such as WASL (Wiskott-Aldrich syndrome like) (Wei et al., 2012) and other WAS-associated proteins like WASF1, WASF2, WAS and WIPF1. Wiskott-Aldrich syndrome (WAS) is a rare X-linked primary immunodeficiency characterized by microthrombocytopenia, eczema, recurrent infections, and an increased incidence of autoimmunity and malignancies (Massaad et al., 2013). FGF10 induces migration and invasion in pancreatic cancer cells through interaction with FGFR2, resulting in a poor prognosis, thus FGF10/FGFR2 signaling is a promising target for new molecular therapy against pancreatic cancer (Nomura et al., 2008).

The main biological processes derived by BioInfoMiner (**Supplementary Table 3**) include Arp2/3 complex-mediated actin nucleation and actin polymerization, involved in multigeneration of dendritic protrusions for 3-dimensional cancer cell migration (Giri et al., 2013).

Prioritized Reactome pathways are shown in **Supplementary Table 4**. The results highlight activation of RHO GTPases, which results in formation of actin stress fibers, lamellipodia and filopodia through interaction with members of the Wiskott-Aldrich Syndrome Protein (WASP) (Vega and Ridley, 2008). In addition, the results indicate increased FGFR signaling, the inhibition of which achieved significant anti-cancer effects in pancreatic cancer (Zhang et al., 2014). Fukushima et al. (2015) showed that loss of free fatty acid receptor FFAR1 in pancreatic cancer cells promoted migration.

These results constitute an accurate and comprehensive interpretation of the underlying molecular complexity, describing the landscape of the molecular interactions captured by the set of mimicked proteins, derived from the Phage Display experiment. **Supplementary Figures 5**, **6** show the extracted networks, through projection on GO Biological Processes and Reactome Pathway hierarchical structures, respectively. The prioritized genes (**Table 1** and **Supplementary Tables 1**, **2**) constitute master regulators based on the topology of these networks, as they have a pivotal role in mediating the cross

TABLE 1 | Summary of the genes prioritized by BioInfoMiner (GO and Reactome).

Gene symbol	GO ranking	Reactome ranking	Description
WASL	1	6	Wiskott-Aldrich syndrome like
FGF10	2	13	fibroblast growth factor 10
WASF2	4	11	WAS protein family member 2
FGF4	8	9	fibroblast growth factor 4
FFAR2	15	4	free fatty acid receptor 2
NKX6-1	3	19	NK6 homeobox 1
WASF1	12	10	WAS protein family member 1
FFAR1	10	17	free fatty acid receptor 1
CLDN17		1	claudin 17
GRSF1		2	G-rich RNA sequence binding factor 1
RPL4		3	ribosomal protein L4
APLN	5		apelin
CD247		5	CD247 molecule
CCL19	6		C-C motif chemokine ligand 19
FBLIM1		7	filamin binding LIM protein 1
HPN	7		hepsin
RPS16		8	ribosomal protein S16
CEBPB	9		CCAAT/enhancer binding protein beta
FMN1	11		formin 1
CLDN2		12	claudin 2
IST1	13		IST1 ESCRT-III associated factor
RPS4Y1		14	ribosomal protein S4 Y-linked 1
VPS4B	14		vacuolar protein sorting 4 homolog B
RPS4X		15	ribosomal protein S4 X-linked
FGF5		16	fibroblast growth factor 5
SOCS7	16		suppressor of cytokine signaling 7
ACVR2A	17		activin A receptor type 2A
EDN1	18		endothelin 1
WAS		18	Wiskott–Aldrich syndrome
FGB	19		fibrinogen beta chain
ANG		20	angiogenin
PDE4D	20		phosphodiesterase 4D
IRS4		21	insulin receptor substrate 4
ZYX	21		zyxin
HOXB2	22		homeobox B2
KRT14		22	keratin 14
WIPF1		23	WAS/WASL interacting protein family member 1

In columns 2 and 3 the rank of each gene according to the GO terms and the Reactome pathways, respectively, is presented and color coded using a gradient from red to yellow. The table is sorted for increasing sum of the two ranks. ***Supplementary Tables 1***, ***2*** *correspond to the original outputs of BioInfoMiner.*

talking between distinct biological processes. This feature is illustrated in **Supplementary Figure 7**, which shows a more compact view of the projection of mimicked proteins on the GO corpus. Systemic processes were derived from semantic clustering of the enriched terms. The prioritized genes are regulators of distinct key processes underlying the PDAC pathology, such as Arp2/3 complex-mediated actin nucleation, Rho protein signal transduction, endothelial cell proliferation, endosome organization, fatty acid signaling and lipopolysaccharide signaling, neutrophil chemotaxis and microtubule polymerization. The oligopeptides mimicking the prioritized protein products can be easily retrieved through the Galaxy interface for further evaluation.

Overall, the present study showcases the capability of the integrative workflow for derivation and selection of biologically active peptides from complex Phage Display experiments, through effective filtering and comprehensive mapping of peptide repertoires on ontological networks and pathways.

DISCUSSION

Phage display coupled with NGS has been introduced almost 10 years ago, thus changing the way of how selected phage repertoires are perceived and analyzed. Deep sequencing techniques provide a global characterization of phage display libraries and selected repertoires, increasing the resolution depth and the potential of the phage display technology for the discovery of target molecules, through the identification of consensus motifs. Today, even the most complex repertoires of selected peptides, usually obtained by *in vivo* selections, can be sampled to obtain a detailed view of their composition and to monitor the progress of the enrichment of specific sequences after each selection/amplification cycle. NGS facilities are easily accessible by the experimentalist and can cover all the steps from the amplification of the DNA of the phages to the delivery of the raw reads in fastq format.

However, the development of analytical tools and strategies is far less advanced. Most software solutions have been in-house developed and not made widely available, by using standard, generic sequence comparison techniques, such as BLAST, or they had a limited scope such as PhastPep (Brinton et al., 2016) taking into account only identical sequences to compare repertoires.

Computational tools for Phage Display data analysis include RELIC (Mandava et al., 2004), PEPTIDE (Pizzi et al., 1995), SiteLight (Halperin et al., 2003), and SLiMFinder (Edwards et al., 2007), which enable sequence alignment and motif detection. However, these tools were designed for small-scale analyses, whereas deeper characterization emerged as a necessity with the advent of NGS techniques. For this purpose, newer methodologies have been developed for high throughput data processing and detection of consensus sequences (Fowler et al., 2011; Alam et al., 2015; Reich et al., 2015), although these techniques did not address the issue of selectivity and comparison among different physiological conditions. This particular problem was addressed by PHASTpep (Brinton et al., 2016), a MATLAB-based tool, which enables differential analysis and selection of peptides that discriminate among different cellular states. PepSimili combines selectivity and network-based functional analysis for prioritization of targets and derivation of biomarkers.

In addition, little has been done to help identifying the spectrum of proteins that are potentially mimicked by the plethora of selected peptides, and to aid in the elucidation of the biological circuits, on which the selection is made. Such information is particularly interesting for biopanning performed

in tissues or *in vivo*, where the complexity of the obtained repertoires reflects the complexity of the biological process under investigation.

In this work, we present a novel strategy based on the identification of proteins containing regions similar to selected peptides obtained by phage display screening. Through their binding to their targets, these peptides are supposed to mimic functional domains of the identified proteins and protein regions. Indeed, most phage displayed peptide selections are originally intended to be used to analyze the repertoires resulting from screenings in complex systems, at least as complex as cell cultures. The presented strategy integrates furthermore the analysis of the retained proteins into a biologically meaningful signal. In contrast to other approaches, PepSimili does not take into account, in the first steps of analysis, the abundance of individual peptides in the repertoires. This is due to the fact that this metric can be greatly affected by the bias of preferential replication of a phage during amplification and the abundance of the target in the system under study, both factors that minimize the interest of its use. However, the abundance of each peptide is reported for mere convenience, in the final results.

Computational derivation of a set of proteins, with domains mimicked by the peptides, can also be helpful for the identification of the targets against which the peptides were selected, while studying their natural binders, as described in interactome databases. In complex systems, for which the peptides or protein domains were identified by our analysis, with the intention to be used as targeting tools for the homing of either therapeutic molecules or imaging agents, it is important to exclude those reacting with targets on healthy tissues. Usually, experimental strategies include the selection of peptides in a system as close as possible to the Test system, to produce a Control repertoire. In the example presented here, the Test system being endothelial cells cultured in tumor conditioned medium, is compared to the Control system being the same cells cultured in fresh medium. Such experimental design favors the identification of targeting molecules that would be specific of the Test system and absent from the Control. Obviously *in vivo* selections are preferable, in order to take into account all potential binding sites of the target molecules, being however far from trivial to be performed, in many cases.

Any set of random peptides presents similarities with the peptides of a set of proteins. It is, therefore, important to be able to evaluate this background noise, which sets an important informational bias confounding the interpretation, and subtract it from the signals obtained by a set of selected peptides, as shown in **Supplementary Figure 4**. In PepSimili, this background noise is systematically subtracted from the signal obtained by the Test repertoire. The approach is general enough and applicable to other high-throughput systems that generate massive peptide repertoires and thus necessitate systematic evaluation and elimination of the background noise. For instance, recently was reported an integrated bacterial system for the discovery of chemical rescuers of disease-associated protein misfolding, which enables massive screening of cyclic oligopeptides with potential pharmacological action against neurodegenerative diseases (Matis et al., 2017). In this system, large combinatorial libraries are biosynthesized in E. coli cells and simultaneously screened for their ability to rescue pathogenic protein misfolding and aggregation, using an ultrahigh-throughput fluorescence-based genetic assay. The high-throughput assay can generate combinatorial libraries of up to 10^8 random peptide sequences. Eventually, coupled with deep sequencing for acquisition of the expressed sequences and *in vitro* validation (Matis et al., 2017), the system derives repertoires of potentially bioactive peptides orders of magnitude smaller (10^2–10^4). However, further *in silico* screening of the oligopeptide repertoires using PepSimili may, on one hand, dramatically reduce the number of candidate oligopeptides, and on the other hand provide a rational basis for peptide selection based on their functional interpretation in terms of impacted biological mechanisms.

Importantly, our methodology enables a systems-level interpretation, through streamlined mapping of the selected mimicked proteins to ontological and pathway networks, providing actionable insights. The BioInfoMiner module derives a small set of orthogonal, systemic processes, accompanied by the respective master regulatory genes linked with a significant part of them, altogether constituting a biomarker signature with actionable potential for clinical, therapeutic or diagnostic processes. Our study demonstrates the efficacy of this integrative workflow using public Phage Display data. Indeed, the tool automatically derived and prioritized key regulators and systemic processes underlying the PDAC pathology. Another potential application area of our approach is the field of Metagenomics, where computational platforms are being constantly developed for analysis, management and annotation of large-scale sequence data (Kunin et al., 2008; Lugli et al., 2016; Koutsandreas et al., 2019). Metagenomic analyses generate massive sequence data, including large amounts of partial or incomplete peptide sequences, stressing the necessity for more efficient annotation methodologies. Our approach that combines massive mapping of peptides to functional networks may enable a more efficient interpretation of the genomic information in the metagenomics content.

Finally, PepSimili is presented in a user-friendly environment, Galaxy, as an integrated tool that performs a complete analysis at a push of a button. A collection of satellite scripts and workflows is also provided, to propose tentative discovery paths that can be followed to complement the actual results, intending to encourage power users to develop, and share with the community new tools/scripts, adding to our work. This implementation facilitates the development of future extensions of the workflow and, importantly, the adaptation of the methodology to other high-throughput technologies, as mentioned above.

AUTHOR CONTRIBUTIONS

AV and KP contributed to PepSimili conception. EP and AC contributed to BioInfoMiner conception and implementation. EP contributed to PepSimili integration. AV, EP, AC, and KP wrote and revised the manuscript.

FUNDING

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Reinforcement of Postdoctoral Researchers" (MIS-5001552), implemented by the State Scholarships Foundation (IKY). This work was funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation 2014–2020" (co-funded by the European Regional Development Fund) and managed by the General Secretariat of Research and Technology, Ministry of Education, Research and Religious Affairs, under the project "Innovative Nanopharmaceuticals: Targeting Breast Cancer Stem Cells by a Novel Combination of Epigenetic and Anticancer Drugs with Gene Therapy (INNOCENT) (MIS 5017608)" (7th Joint Translational Call – 2016, European Innovative Research and Technological Development Projects in Nanomedicine) of the ERA-NET EuroNanoMed II. This work was also supported by INSERM and the French National Research Agency (ANR 2006–2015). All supports are gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2019.01160/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Alam, K. K., Chang, J. L., and Burke, D. H. (2015). FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids* 4:e230. doi: 10.1038/mtna.2015.4

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bábíčková, J., Tóthová, Ľ, Boor, P., and Celec, P. (2013). *In vivo* phage display — a discovery tool in molecular biomedicine. *Biotechnol. Adv.* 31, 1247–1259. doi: 10.1016/j.biotechadv.2013.04.004

Brinton, L. T., Bauknight, D. K., Dasa, S. S. K., and Kelly, K. A. (2016). PHASTpep: analysis software for discovery of cell-selective peptides via phage display and next-generation sequencing. *PLoS One* 11:e0155244. doi: 10.1371/journal.pone.0155244

Chatziioannou, A., and Moulos, P. (2011). Exploiting statistical methodologies and controlled vocabularies for prioritized functional analysis of genomic experiments: the StRAnGER web application. *Front. Neurosci.* 5:8. doi: 10.3389/fnins.2011.00008

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucl. Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Deutscher, S. L. (2010). Phage display in molecular imaging and diagnosis of cancer. *Chem. Rev.* 110, 3196–3211. doi: 10.1021/cr900317f

Dias-Neto, E., Nunes, D. N., Giordano, R. J., Sun, J., Botz, G. H., Yang, K., et al. (2009). Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS One* 4:e8338. doi: 10.1371/journal.pone.0008338

Edwards, R. J., Davey, N. E., and Shields, D. C. (2007). SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* 2:e967. doi: 10.1371/journal.pone.0000967

Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011). Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* 27, 3430–3431. doi: 10.1093/bioinformatics/btr577

Fukushima, K., Yamasaki, E., Ishii, S., Tomimatsu, A., Takahashi, K., Hirane, M., et al. (2015). Different roles of GPR120 and GPR40 in the acquisition of malignant properties in pancreatic cancer cells. *Biochem. Biophys. Res. Commun.* 465, 512–515. doi: 10.1016/j.bbrc.2015.08.050

Giri, A., Bajpai, S., Trenton, N., Jayatilaka, H., Longmore, G. D., and Wirtz, D. (2013). The Arp2/3 complex mediates multigeneration dendritic protrusions for efficient 3-dimensional cancer cell migration. *FASEB J.* 27, 4089–4099. doi: 10.1096/fj.12-224352

Halperin, I., Wolfson, H., and Nussinov, R. (2003). SiteLight: binding-site prediction using phage display libraries. *Protein Sci.* 12, 1344–1359. doi: 10.1110/ps.0237103

Hume, T., Soueidan, H., Wong, J. T., Vekris, A., Petry, K. G., Nikolski, M., et al. (2013). Efficient mapping of short peptides on whole proteome database for biomarker discovery. in *Proceedings of the Conference on JOBIM in 2013*, France.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Kolonin, M. G., Sun, J., Do, K.-A., Vidal, C. I., Ji, Y., Baggerly, K. A., et al. (2006). Synchronous selection of homing peptides for multiple tissues by *in vivo* phage display. *FASEB J.* 20, 979–981. doi: 10.1096/fj.05-5186fje

Koutsandreas, T., Binenbaum, I., Pilalis, E., Valavanis, I., Papadodima, O., and Chatziioannou, A. (2016). Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks. *Int. J. Monitor. Surve. Technol. Res.* 4, 30–49. doi: 10.4018/IJMSTR.2016040103

Koutsandreas, T., Ladoukakis, E., Pilalis, E., Zarafeta, D., Kolisis, F. N., Skretas, G., et al. (2019). ANASTASIA: an automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data. *Front. Genet.* 10:469. doi: 10.3389/fgene.2019.00469

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578.

Lugli, G. A., Milani, C., Mancabelli, L., van Sinderen, D., and Ventura, M. (2016). MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol. Lett.* 363:fnw049. doi: 10.1093/femsle/fnw049

Mandava, S., Makowski, L., Devarapalli, S., Uzubell, J., and Rodi, D. J. (2004). RELIC– A bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites. *Proteomics* 4, 1439–1460. doi: 10.1002/pmic.200300680

Massaad, M. J., Ramesh, N., and Geha, R. S. (2013). Wiskott-Aldrich syndrome: a comprehensive review: wiskott-aldrich syndrome. *Ann. N.Y. Acad. Sci.* 1285, 26–43. doi: 10.1111/nyas.12049

Matis, I., Delivoria, D. C., Mavroidi, B., Papaevgeniou, N., Panoutsou, S., Bellou, S., et al. (2017). An integrated bacterial system for the discovery of chemical rescuers of disease-associated protein misfolding. *Nat. Biomed. Eng.* 1, 838–852. doi: 10.1038/s41551-017-0144-143

Moutselos, K., Maglogiannis, I., and Chatziioannou, A. (2011). "GOrevenge: a novel generic 1reverse engineering method for the identification of critical molecular players, through the use of ontologies," in *Proceedings of the IEEE Transactions on Biomedical Engineering*, Piscataway, NJ.

Nomura, S., Yoshitomi, H., Takano, S., Shida, T., Kobayashi, S., Ohtsuka, M., et al. (2008). FGF10/FGFR2 signal induces cell migration and invasion in pancreatic cancer. *Br. J. Cancer* 99, 305–313. doi: 10.1038/sj.bjc.6604473

Pasqualini, R., and Ruoslahti, E. (1996). Organ targeting *In vivo* using phage display peptide libraries. *Nature* 380, 364–366. doi: 10.1038/380364a0

Pilalis, E., Koutsandreas, T., Valavanis, I., Athanasiadis, E., Spyrou, G., and Chatziioannou, A. (2015). KENeV: a web-application for the automated

reconstruction and visualization of the enriched metabolic and signaling superpathways deriving from genomic experiments. *Comp. Struc. Biotechnol. J.* 13, 248–255. doi: 10.1016/j.csbj.2015.03.009

Pilalis, E. D., and Chatziioannou, A. A. (2013). "Prioritized functional analysis of biological experiments using resampling and noise control methodologies," in *Proceedings of the 13th IEEE International Conference on BioInformatics and BioEngineering*, (Chania: IEEE).

Pizzi, E., Cortese, R., and Tramontane, A. (1995). Mapping epitopes on protein surfaces. *Biopolymers* 36, 675–680. doi: 10.1002/bip.360360513

Reich, L. L., Dutta, S., and Keating, A. E. (2015). SORTCERY–a high–throughput method to affinity rank peptide ligands. *J. Mol. Biol.* 427, 2135–2150. doi: 10.1016/j.jmb.2014.09.025

Rodi, D. J., and Makowski, L. (1999). Phage-display technology – finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.* 10, 87–93. doi: 10.1016/S0958-1669(99)80016-80010

Schwenk, J. M., Omenn, G. S., Sun, Z., Campbell, D. S., Baker, M. S., Overall, C. M., et al. (2017). The human plasma proteome draft of 2017: building on the human plasma peptideatlas from mass spectrometry and complementary assays. *J. Proteome Res.* 16, 4299–4310. doi: 10.1021/acs.jproteome.7b00467

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43, W589–W598. doi: 10.1093/nar/gkv350

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197. doi: 10.1016/0022-2836(81)90087-90085

Vega, F. M., and Ridley, A. J. (2008). Rho GTPases in cancer cell biology. *FEBS Lett.* 582, 2093–2101. doi: 10.1016/j.febslet.2008.04.039

Wei, P., Tang, H., and Li, D. (2012). Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PLoS One* 7:e46887. doi: 10.1371/journal.pone.0046887

Zhang, H., Hylander, B. L., LeVea, C., Repasky, E. A., Straubinger, R. M., Adjei, A. A., et al. (2014). Enhanced FGFR signalling predisposes pancreatic cancer to the effect of a potent FGFR inhibitor in preclinical models. *Br. J. Cancer* 110, 320–329. doi: 10.1038/bjc.2013.754

Conflict of Interest: EP and AC were employed by company eNIOS Applications P.C.

The remaining 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.

The reviewer MV and the handling Editor declared their shared affiliation at the time of the review.

Copyright © 2019 Vekris, Pilalis, Chatziioannou and Petry. 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) and the copyright owner(s) 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 Genetics

TECHNOLOGY REPORT
published: 24 May 2019
doi: 10.3389/fgene.2019.00469

ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next Generation Sequencing Data

Theodoros Koutsandreas[1,2†], Efthymios Ladoukakis[1,3†], Eleftherios Pilalis[1,2], Dimitra Zarafeta[1], Fragiskos N. Kolisis[1,3], Georgios Skretas[1] and Aristotelis A. Chatziioannou[1,2]*

[1] Institute of Chemical Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece, [2] e-NIOS Applications PC, Athens, Greece, [3] Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Athens, Greece

OPEN ACCESS

Edited by:
Marco Antoniotti, University of Milano-Bicocca, Italy

Reviewed by:
Cuncong Zhong, University of Kansas, United States
Digvijay Verma, Babasaheb Bhimrao Ambedkar University, India

***Correspondence:**
Aristotelis A. Chatziioannou
achatzi@eie.gr

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 November 2018*
Accepted: *01 May 2019*
Published: *24 May 2019*

Citation:
Koutsandreas T, Ladoukakis E, Pilalis E, Zarafeta D, Kolisis FN, Skretas G and Chatziioannou AA (2019) ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next Generation Sequencing Data. Front. Genet. 10:469. doi: 10.3389/fgene.2019.00469

Metagenomic analysis of environmental samples provides deep insight into the enzymatic mixture of the corresponding niches, capable of revealing peptide sequences with novel functional properties exploiting the high performance of next-generation sequencing (NGS) technologies. At the same time due to their ever increasing complexity, there is a compelling need for ever larger computational configurations to ensure proper bioinformatic analysis, and fine annotation. With the aiming to address the challenges of such an endeavor, we have developed a novel web-based application named ANASTASIA (automated nucleotide aminoacid sequences translational plAtform for systemic interpretation and analysis). ANASTASIA provides a rich environment of bioinformatic tools, either publicly available or novel, proprietary algorithms, integrated within numerous automated algorithmic workflows, and which enables versatile data processing tasks for (meta)genomic sequence datasets. ANASTASIA was initially developed in the framework of the European FP7 project HotZyme, whose aim was to perform exhaustive analysis of metagenomes derived from thermal springs around the globe and to discover new enzymes of industrial interest. ANASTASIA has evolved to become a stable and extensible environment for diversified, metagenomic, functional analyses for a range of applications overarching industrial biotechnology to biomedicine, within the frames of the ELIXIR-GR project. As a showcase, we report the successful *in silico* mining of a novel thermostable esterase termed "EstDZ4" from a metagenomic sample collected from a hot spring located in Krisuvik, Iceland.

Keywords: metagenomics, bioinformatics, next generation sequencing, automated pipelines, systemic biology, novel enzymes

INTRODUCTION

DNA sequencing techniques have advanced at a prodigious rate during the last decade, attaining higher yields in conjunction with minimizing costs per sequencing run (Stein, 2010). This culminates with the advent of next-generation sequencing (NGS) comprising rapid, high-throughput protocols that generate large amounts of high quality data of deep coverage for only

a small fraction of the cost of traditional sequencing technologies (i.e., Sanger). The continuous advancement of NGS techniques has resulted in a subsequent rapid progress in the field of metagenomics, revolutionizing the methodologies for in-depth exploration of the genomic and subsequent enzymatic content of microbial communities without the need of prior culturing. This development, however, is bringing up numerous challenges in the processing of the resulting raw data, thus rendering the bioinformatic analysis a major bottleneck, in any metagenomics project (Scholz et al., 2012). The arising bioinformatic challenges originate from the massive amount of raw data that NGS techniques generate, but also from the vast diversity of bioinformatic tools essential for all gene detection, and annotation tasks that constitute a full metagenomic analysis pipeline (Kunin et al., 2008). In addition, the immense demand on computational and storage resources for such an analysis practically imposes the use of cloud computing methodologies, so as to ensure its feasibility in real world datasets (Wilkening et al., 2009). To address these issues we have developed ANASTASIA (automated nucleotide aminoacid sequences translational plAtform for systemic interpretation and analysis), a user-friendly, web-based, computational infrastructure, which comprises numerous bioinformatic tools, integrated as modular components of automated workflows, and which has been explicitly designed to handle a wide range of metagenomic analyses. This framework is an extended instance of the Galaxy platform (Goecks et al., 2010), customized specifically for metagenomics-related bioinformatic analyses and is linked to a MariaDB database (The MariaDB Foundation, 2009) server via a Web2py (Pierro, 2010) web interface, in order to facilitate the efficient management of the resulting data files. This framework retains all the features of a classic Galaxy instance thus offering the option of being a portable solution while enabling its further customization by integrating additional tools and workflows. ANASTASIA has been tested extensively in diverse metagenomics datasets, as it was initially developed to be the core bioinformatics tool of the EU FP7 project HotZyme[1], which targeted to the exhaustive analysis of metagenomes from thermal springs (Wohlgemuth et al., 2018), with the scope of evolving into a stable environment for powerful, and bionformatic analyses that may support a broad range of metagenomics analysis tasks. ANASTASIA is certainly the first of its kind platform, with this level of operational maturity and tested stability, to provide all-inclusive workflows, handling each part of the long, and diverse list of steps comprising an exhaustive analysis of metagenomic data, being at the same time, fully customizable by its users. In contrast to the limited analytic options and limited automation offered by other pipelines with the same broader scope, ANASTASIA's automated pipelines are able to process various computational steps, from handling raw sequencing data to assigning putative function predictions for gene encoding sequences, or providing powerful functional characterization of the underlying emerging molecular networks. Based on all the aforementioned, it is capable of addressing different problems, like the screening of thermophiles or the systematic screening of the human microbiome in various infections, as part of the Hellenic Bioinformatic computational Infrastructure ELIXIR-GR, which represents the Hellenic node of ELIXIR.

MATERIALS AND METHODS

Design and Implementation

The backbone of the ANASTASIA platform was assigned to a server by the name "Motherbox" owned by the National Technical University of Athens (NTUA) that hosted a local instance of the Galaxy platform. The Motherbox server is equipped with 64 CPU cores, 512 GB RAM and a total of 7.2 TB disk capacity. The Galaxy installation was performed by downloading the latest version of the source code from the Mercurial (Mercurial, 2005) depository[2] and running its startup script (run.sh) in order to automatically download all dependent Python (The Python Programming Language, 2001) modules ("eggs") that are essential. As part of the installation process, the local MariaDB database server was exploited and installed on Motherbox and linked to the Galaxy server through integration of the appropriate custom initialization arguments in the system configuration file (galaxy.ini) of the platform. The web accessibility of Galaxy was secured by configuring Motherbox's Apache web server (The Apache, 1995) to proxy any requests to the virtual host motherbox.chemeng.ntua.gr/anastasia_dev/ to a dedicated port on the server, thus broadcasting the platform to all external users via the aforementioned web address. Extra customization to the configuration files of the Galaxy instantiation to include and specify the bioinformatic tools and algorithms available after integration in order to become embedded in ANASTASIA's front-end list of tools.

Front-End Customization

The ANASTASIA front-end was developed by altering the graphical user interface (GUI) of the Galaxy platform into a more intuitive and user-friendly layout. The new GUI, was based on scripts written in JavaScript (Javascript, 2016) that modify the typical settings of the homepage layout of the Galaxy platform and that add several different menu and submenu buttons, which correspond to distinct categories of the integrated bioinformatic tools and available workflows. Most of the tools that were incorporated into the original Galaxy instance were removed and have been replaced by appropriately curated and tested tools designed for metagenomic analytical tasks, such as *de novo* sequence assembly, open reading frame (ORF) detection, homology based protein prediction, machine-learning-based protein prediction, etc. This was accomplished by developing tuned parser algorithm scripts in Perl (The Perl Programming Language, 2002) and Python language, as well as by preparing appropriate configuration files in Extensible Markup Language (XML) (Extensible Markup Language, 2013) for each individual tool. In that way, the integration of each tool was managed by means of a Galaxy-generated GUI (written in

[1] hotzyme.com

[2] https://bitbucket.org/galaxy/galaxy-dist/

XML) that links the necessary input parameters into the parser script which, in turn, invokes the corresponding executable. The choice of tools that were installed on the server and integrated into ANASTASIA was based mainly on three criteria: (1) their overall performance vs. their computational cost plus resource demands, (2) their compatibility with the Galaxy platform, and (3) their user-friendliness (i.e., advanced visualization functionality, interactive data management in local computers, such as MEGAN). While the first criterion is important for tasks, such as contig assembly and sequence similarity searches, which constitute the computational bottlenecks of every bioinformatic analysis of sequencing data, the latter two were the main consideration for less costly parts of the analysis, but of great importance regarding the quality of the interpretation. In addition to already established bioinformatic tools, the novel interface of ANASTASIA was enriched with in-house algorithms that enable numerous data processing tasks such as functional analysis and data management among others (see below).

Data Acquisition

A tool for direct file uploading to the platform was included in the original Galaxy instance and was kept available as it enables both access to local data from a user's personal computer, as well as data available online via URL. In addition, numerous sequencing raw data files were linked to ANASTASIA via their directory path in the Motherbox server e.g., large sequencing datasets from the HotZyme project. This enables their direct utilization by each user precluding the creation of additional copies for each analysis. This became feasible by altering Galaxy's system configuration files and may be used for any dataset import on the server, provided the availability of its directory path.

Sequencing Quality Control

Sequencing quality is expressed based on Phred base calling algorithm (Ewing et al., 1998), which assigns a quality score Q to each base, proportional to its error probability P. Each score is calculated using the formula: $Q = -10\log_{10}P$. The data from any sequencing experiment generate a second file containing the scores matching the sequencing reads from the first one or, more commonly, combine the two in the same file, i.e., FASTQ format. Any sequencing quality control analysis requires a tool to parse through the sequencing data files and according to the user's specifications either trim or exclude highly problematic reads. For such a task there are various bioinformatic solutions (Davis et al., 1998; FastQC, 2010; Fastx Toolkit, 2009; Pandey et al., 2016), all of which manage to handle the issue quite rapidly, and memory efficiently (Davis et al., 1998). In order to enable sequencing quality control in ANASTASIA and considering the somewhat similar performance of all the above-mentioned tools, the FASTX toolkit (Fastx Toolkit, 2009) and FastQC (FastQC, 2010) were chosen for integration as they were already included in the original Galaxy instance. These tools were installed on Motherbox server and their corresponding Galaxy parsers and XML configuration files were preserved in the final platform.

Taxonomic Analysis

There are two major identification schemes for microbial species in a metagenomic sample, the first being based on amplicon sequencing (e.g., 16S rRNA), while the latter on whole metagenome sequencing (Garrity, 2016). Since this first version of ANASTASIA is currently dedicated to analyzing whole metagenome sequencing datasets and as this approach for taxonomic classification has been proven more efficient than amplicon sequencing (Escobar-Zepeda et al., 2018), the appropriate tools had to be considered for integration. The MEGAN (Huson et al., 2007) software was chosen, due to its advanced operational features, including both Linux and Windows support, providing an interactive GUI for further result visualization and interactive data management on a personal computer, as well as its potential to be used as a metabolic pathway analysis tool, exploiting reference databases such as KEGG (Kanehisa et al., 2017), SEED (Overbeek et al., 2005), and COG (Tatusov et al., 2000). The tool operates by receiving the resulting dataset from a sequence similarity search analysis such BLAST as input, in order to calculate the percentage of each species present in the sample while yielding summarized results as visual representations (e.g., bar charts).

De novo Sequence Assembly

One of the major bottlenecks, in terms of execution times and computational demands, is the assembly of reads into contiguous sequences (contigs), mostly due to the ever higher throughputs of NGS technologies, of small read lengths. Consequently, the sequence assembly methodologies transition from the overlap-layout-consensus (OLC) algorithms toward De-Bruijn-graph based paradigms (Li et al., 2012). The choice of the proper assembler relies on two equally important points: (1) the quality of the produced assembly and (2) the required computational resources. Numerous studies have managed to extensively compare various assemblers either for metagenomic data (van der Walt et al., 2017; Vollmers et al., 2017) or for single organism genomic data (Earl et al., 2011; Bradnam et al., 2013) and while there seems to be a slight advantage of metaSPAdes (Nurk et al., 2017), in terms of length of the produced contigs, the final choice for ANASTASIA was Megahit (Li et al., 2015) as it attained a comparable performance with much smaller computational resources than the typical configuration of metaSPAdes. The Velvet (Zerbino and Birney, 2008) assembler was also included in the final instance because of its popularity, as it provides visualization options for its assembly via the generation of afg files and is already included in the repository for Galaxy tool integration Galaxy Tool Shed[3]. Both tools were installed on the Motherbox server and later linked to the platform via the appropriate changes in Galaxy's tool configuration file (tool_conf.xml). Megahit was configured manually via an in-house XML script integrating it to ANASTASIA.

ORF/Gene Detection

De novo gene detection, while not being significantly challenging in terms of computational cost, is the most crucial part

[3]http://toolshed.g2.bx.psu.edu/

in any analysis oriented toward unearthing novel enzymes from metagenomic samples. From an algorithmic perspective, detecting a gene in a large contiguous sequence could theoretically require only the extraction of the appropriate ORF as performed by the Getorf tool of the EMBOSS suite (Rice et al., 2000). Nevertheless the existence of an ORF cannot guarantee that the specific genomic region is translated as is. For example, spurious ORFs (Veloso et al., 2005), especially in high GC content genomes, can often lead to the detection of numerous false potential gene sequences within the same region. Moreover, sequence assembly can sometimes be omitted entirely or, depending on parameter settings, read coverage and abundance of species, result in shorter contigs that correspond to incomplete genomes thus rendering potential ORFs undetectable. Various algorithms (Noguchi et al., 2006, 2008; Rho et al., 2010; Zhu et al., 2010) have emerged that are dedicated to working with short sequences (either shorter contigs or even reads) but as an extensive third-party benchmarking study is yet to be published, FragGeneScan tool was the choice for integration with ANASTASIA as it seems to outperform all its predecessors (Rho et al., 2010). This tool was installed on the Motherbox server but its wrapper script (run_FragGeneScan.pl) had to be edited, in the way it produced the output files, in order to be more compatible with a Galaxy instance that would host numerous users simultaneously. The approriate XML configuration files for integration with ANASTASIA were imported by Galaxy Tool Shed and edited appropriately in order to comply with the abovementioned changes in the wrapper script.

Protein Function Prediction

In order to predict the putative function of the sequences retrieved from the ORF/gene detection analysis of the BLAST suite (Altschul et al., 1990) alignment tools and the HMMER (Finn et al., 2011) software were installed on the server and integrated on ANASTASIA by using the publicly available XML configuration files found in Galaxy's Tool Shed (Galaxy Tool Shed, 2005). For the BLASTn and BLASTp tools, which have been designed to handle nucleotide and amino-acid sequences respectively, the NCBI-nt, NCBI-nr and UniProt databases were downloaded on the server and were parsed with the appropriate BLAST commands (makeblastdb) for immediate use. Moreover, for the purposes of HotZyme project, code was written in Perl that enabled the creation and formatting of a BLAST-oriented, customizable database, which contained all of the annotated sequences with hydrolytic activities derived from the SwissProt (Apweiler et al., 2004) database. For the HMMER tool, the Pfam (Sonnhammer et al., 1997) database was downloaded and formatted for use with the appropriate commands (hmmpress). To enrich ANASTASIA's protein prediction capabilities, we applied machine-learning based methodologies that enable us to translate the genomic content of any environmental sample. These include the installation and integration into ANASTASIA of the EFICAz (Kumar and Skolnick, 2012) software and of the PROKKA (Seemann, 2014) pipeline, which, in spite of the numerous programs, it calls upon to perform a complete analysis, is introduced to the user as a single tool.

Data Management/ANASTASIA Knowledgebase Design

To avoid transactional locks generated with the increase of the platform's users and to improve overall efficiency, ANASTASIA was linked to a local MariaDB database on Motherbox by editing the Galaxy setup configuration file (galaxy.ini). In order to facilitate the parsing of resulting data (e.g., contig assembly or BLAST result files), in-house algorithms written in Python were integrated as tools in the platform. Each Python parser imports the corresponding data into the MariaDB database on the Motherbox server and returns it as a dump file allowing the user to download it and re-import it into their local MariaDB or MySQL (MySQL, 1995) database server. Additionally, the parsers designed to handle BLAST results, return FASTA files with all the sequences from the original dataset that do not return any alignment hit above the predetermined statistical thresholds. For the purposes of online inspection and user interaction with the scope of filtering analysis results, the ANASTASIA Knowledgebase was developed, a MariaDB database system that allows users to import their results (as MariaDB/MySQL dump files) and access them via a Web2py interface. This was implemented through an in-house tool developed in Python and integrated into ANASTASIA. This tool creates a new database schema, where it imports the data and appends the new schema information in Web2py's "controller" Python scripts. In this way the schema is afforded via a user interface, which also contains all MariaDB query search capabilities for quick and efficient data manipulation, filtering, and retrieval. The created web interface is secured via an access control system, which the user has to apply for in order to get the necessary username and password credentials from an administrator. The tool generates a unique id for each database entry, available only to the user, with which the data can be retrieved at any later time. Each user can access the knowledgebase either directly from ANASTASIA's front end or from the link generated from the tool that includes also the unique id for their datasets.

Supplementary Tools

Additional tools, based on both published and in-house scripts, have also been integrated in the platform and perform various non-trivial tasks such as sequence clustering (Li and Godzik, 2006), sequence translation (Fastx Toolkit, 2009), and data format management, etc. These tasks have also been incorporated in the subsequent workflows (see *Metagenomic workflows)* in order to format datasets accordingly for optimal utilization in each different part of the analysis by the respective tool(s).

Metagenomic Workflows

The aforementioned tools have been assembled together as modules of automated workflows in ANASTASIA that can handle any type of metagenomic dataset and analysis. These workflows enable their comprised tools to exchange input and output data with each other, in order for each one of them to execute the specific analytical task, automatically, without any user interference e.g., the gene detection tool will automatically use the output data from the assembly tool

and will in turn provide its output data to the BLAST tool for further annotation. This was made possible by exploiting Galaxy's Workflow Canvas, in order to design the sequel of the various analytical processes, or by simply extracting the history of each of our past analysis in the HotZyme project as a complete, and ready to use workflow. These capabilities of the Galaxy application were retained in ANASTASIA so as to inherit the users with the capability to create their own customized workflows on top of the default ones we include. ANASTASIA comes with default automated workflows that allow a complete analysis depending on the input data and are named accordingly. In addition, these workflows might also be fully customizable if imported to a user's account, something which may prove extremely useful and extend their functionality. This can be exploited for instances on machines with limited resources where handling computationally intensive parts of the analysis, e.g., similarity searches via BLAST is essential. The "Starting From FASTQ Reads" workflow is designed for raw sequencing reads in FASTQ format as input and includes the following tasks: (i) quality control by FASTX toolkit; (ii) assembly into contigs by Megahit; (iii) gene identification by FragGeneScan; (iv) gene annotation using a combination of BLAST, PROKKA, HMMER and EFICAZ; (v) visualization of results using in-house parser scripts that import the annotation results into the server's MariaDB database and visualize it via Web2py. The "Starting From FASTA Reads" workflow has the same functionality as the one mentioned above but is designed for FASTA formatted datasets hence omitting the quality control steps. The "After Assembly" workflow follows all the previous annotation steps but starting with contig FASTA files as input i.e., the results of the assembly process. The "Taxonomic and functional analysis" workflow is designed for detecting the different microbial populations in a metagenomic sample and requires as input a FASTA file of raw sequencing reads that subjects to the following analysis: (i) homology analysis using BLAST tool against NCBI-nr database, (ii) taxonomic analysis using MEGAN software. Every workflow consists of tools and algorithms integrated in ANASTASIA with each and every one of them being available as a standalone tool for the user to exploit in customized analyses.

Biotranslator Workflow

Functional enrichment analysis constitutes the foremost approach to interpret the impact of a set of genes (or gene products) to the cellular physiology, namely the co-regulation of distinct cellular mechanisms that gives rise to diverse phenotypes. Conceptually, it is based on the association of genes with semantic terms, which refer to molecular pathways, cellular components, biological mechanisms, or phenotypic traits. Those terms are predominantly organized in logical structures, which describe the knowledge of a specific biological domain (Gross et al., 2016). In order to aid the elucidation of the biological underpinnings of an unknown gene set, each term needs to be annotated with genes, and based on prior inferences of scientific studies. Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa et al., 2017), and reactome pathways (Fabregat et al., 2018) are well-established omnibuses, which correlate systematically their descriptive terms with genes by following a hierarchical structure of deductive steps and constructing the appropriate framework for the functional enrichment analysis. However, as the scientific community is mainly focused on organisms of traditional biomedical interest, they produce electronically, and manually curated annotations only for a limited organismal spectrum, including human, mouse, other model organisms and some specific bacteria, which are a negligible ratio of the whole prokaryotes' kingdom.

Nowadays, various tools and software are used for the functional enrichment analysis of significant gene lists, derived from -omics experiments. Their core computational process consists of over representation of statistical tests and p-value correction approaches, in order to minimize the amount of false positives. The StRAnGER algorithm (Chatziioannou and Moulos, 2011) applies a non-parametric procedure, targeting to mitigate experimental and annotation noise by filtering out any trivial and non-informative term and reveal system-level terms, which reflect the underlying components of the examined phenotype. As a result, it translates the individual genes, through the aforementioned vocabularies, to a restricted list of prioritized terms, which could be depicted as a broad network of functional of phenotypic entities.

This network represents a descriptive snapshot of the biological problem under investigation, incapable to condense the interpretation to few precise markers. GOrevenge (Moutselos et al., 2011) has been developed to surpass that limitation of functional enrichment analysis, aiming to propose potential regulatory hub genes or gene products, or even further putative marker signatures, related to the ranked set of over-represented terms. It uses graph-theoretical methods, exploiting the graphical structure of the controlled databases, such as the direct acyclic graph of GO, to detect cross-linked entities, which take part in many topologically distinct nodes of the graph. In this way, the output of an omic experiment may result to a compact list of prioritized genes or proteins, without the use of prior knowledge, or human supervision (i.e., phenotype related seed sets or keywords), reducing problem dimensionality to a succinct set of features.

As our motivation here was to provide automated solutions for the functional description of microbial communities, the aforementioned approaches sought adaptive adjustments. A common processing of metagenomic raw data correlates the detected ORFs with known sequences or protein domains, based on sequence alignment algorithms. The functional interpretation could rely on those sequence similarity predictions. Besides the debatable assumption that sequence similarity implies or not functional relevance (Gerlt and Babbitt, 2000), the overriding problem is that the existed functional enrichment analysis tools use organism-specific data, in contrast to the fundamental concept of metagenomics studies. Metagenomics explore the microbial communities as a unified entity, endeavoring to detect and decipher the synergistic mechanisms that cause community homeostasis or other biotechnologically interesting features (such as thermostability and resilience to acid environments). As a result, an appropriate functional enrichment analysis tool needs to take into account all the

available functional annotation of the prokaryotic world, combined in a unified database.

UniProt-SwissProt knowledgebase (Apweiler et al., 2004) includes manually curated descriptions of more than 350k known proteins from the prokaryotic world. To overcome the existence of organism-specific databases, such as *Escherichia coli* or Bacillus subtilis, we exploited the whole mapping of UniProt-SwissProt knowledgebase. We combined data from different organisms (bacteria and archaea), related to gene products and GO terms associations and producing a unified schema for GO. All amino acid sequences, which originated from the same gene in different organisms, were conceptually clustered together, combining their functional annotations to produce a unified gene – GO terms mapping. In order to eliminate the annotation bias of extensively studied prokaryotes, infrequent associations were filtered out. The relative frequency of each gene – GO term pair was calculated as the ratio of organisms which include that pair in their mapping to all the organisms, which contain that gene in their DNA. Gene-specific distributions of relative frequencies were constructed so that every pair with value lower than the respective distribution median was excluded from the final annotation schema. The output graphs of GO constitute a global description of biological processes, cellular components and molecular functions that exist in the prokaryotic kingdom and are correlated with at least one gene, independently of taxonomic details. Such an ensemble of annotations could be used for the biological interpretation of microbial communities, regardless their population distribution, and taxonomic profile.

Summarizing, in the framework of the ANASTASIA platform, a new workflow was developed, named BioTranslator, for the functional interpretation of metagenomic data, which encompasses sequential computational steps, and adapted to the particularities of input data. In order to analyze a metagenomic sample, the user is able to import either a BLASTp output (specific tabular format and executed on the SwissProt database) or a list of gene symbols, derived from previous analytical tasks. Targeting the detection of the most trustworthy BLAST hits, BioTranslator adopts strict alignment criteria, filtering out matches with query coverage lower than 90% or subject coverage lower than 50% while taking into consideration a user-defined threshold about the hits' *e*-value. The best hit of each query is kept and UniProt IDs are translated into the respective gene symbols. Regardless of the initial input, StRAnGER performs the functional enrichment analysis of genes list and derives a set of statistically significant terms, as they are described in the three GO domains. The user defines the domain which will be used for the prioritization of genes. Hence, GOrevenge uses the enriched part of that domain in order to exploit its topological characteristics and disclose the most critical genes that could be assumed as the master regulators, bearing a part of causality of community features.

Application of ANASTASIA

Automated nucleotide aminoacid sequences translational plAtform for systemic interpretation and analysis was exhaustively tested in various aspects, during the HotZyme project where it was mainly used to store, manage and annotate metagenomic sequencing data, taken from eight remote hot springs around the world (Menzel et al., 2015), in order to detect novel thermostable enzymes of industrial interest. The first beta version of the platform was installed on a server of the University of Copenhagen named "Helios" and provided access both to the data and to the annotation tools in order to predict sequences that correspond to thermostable enzymes of potential hydrolytic activity which could be verified in the lab at a later point. In this first version, during the development of its various automated workflows and corresponding modules, we exploited each integrated algorithm to run our first analyses of the samples, which, in turn, resulted in the detection of various novel enzymes exhibiting enhanced thermostability (Zarafeta et al., 2016a,b). The final version of ANASTASIA was fed again with pre-existing data for another iteration of the analysis but this time via its automated workflows of fined-tuned algorithms (supplied with default parameters for optimal performance) and resulted in the detection of an additional novel enzyme, termed EstDZ4, as described below.

Identification of the *estDZ4* Gene

The raw sequencing data from each sample of the HotZyme project were imported in the server Helios of the University of Copenhagen and were linked to ANASTASIA, thus making it directly available both for download and analysis via its integrated tools and workflows. Sequencing assembly was performed (Menzel et al., 2015) from University of Copenhagen and resulting contig datasets were also imported in ANASTASIA for further analysis. The selection of the sequence of *estDZ4* as a candidate gene encoding a protein with putative esterolytic activity occurred through the application of the "After assembly" workflow on the assembly data of the sample Is3-13 originating from a high temperature pool (90°C/pH 3.5–4.0) in Krisuvik, Iceland (Menzel et al., 2015). The assembly dataset of the Is3-13 sample was 29.8MB in size and consisted of 34.651 contigs originating from 10,050,000 raw reads. Running time on the Motherbox server (512GB RAM, 64CPUs) for the same workflow, using the above-mentioned dataset, lasted 24 h, 54 min, and 20 s with the major bottlenecks being, as expected, the similarity searches to the locally downloaded databases of NCBI-nr (15 h, 27 min, and 50 s) and Pfam-A (7 h, 27 min, and 34 s). The results were downloaded as sqldump files and examined on MySQL workbench where the sequence for EstDZ4 was chosen for further curation. EstDZ4 was chosen because it exhibited 99% identity (query coverage 98%) to a putative esterase/lipase from *Thiomonas sp.CB3* [GenBank: CQR41430.1] in NCBI-nr but only 23% identity (query coverage 84%) to a previously characterized GDSL esterase/lipase from *Arabidopsis thaliana* [UniProtKB/Swiss-Prot: Q9FIA1.1]. Futhermore, it was predicted to contain a GDSL-like Lipase/Acylhydrolase catalytic domain from Pfam-A database.

The representative (non redundant) putative gene sequences detected from the initial steps [FragGeneScan and CD-HIT (Li and Godzik, 2006)] of the above mentioned pipeline were used as input in the BioTranslator workflow, producing the pathway analysis results in a total of 36 min and 5 s. Once again, the

similarity search analysis (BLASTp against a customized database containing only the prokaryote entries from SwissProt) was the major bottleneck requiring a total 31 min and 37 s.

Cloning, Purification, and Biochemical Characterization of EstDZ4

Construction of the pASK-EstDZ4 plasmid was carried out by amplifying *estDZ4* from the isolated metagenomic DNA by polymerase chain reaction (PCR) using primers containing an XbaI site (5′- AAAAATCTAGAAGGAGGAAACGATGTCCGT GGCGAGTGTGAATTCGGCC-3′) and an XhoI site along with octahistidine tag (5′- AAAAAACTCGAGTTAGTGGTGGT GGTGGTGGTGGTGGTGTTGCGAAATCCAGCCAAAACCC-3′) (restriction sites underlined, octahistidine tag doubly underlined). The forward primer was designed so as not to include amino acids 1–38, which were predicted to correspond to a signal sequence according to HMMER analysis. The PCR product was cloned into the expression vector pASK75 (Skerra, 1994). *E. coli* Origami 2 (DE3) (Novagen) cells were transformed with pASK-EstDZ4, grown in 5 ml of Luria-Bertani (LB) medium containing 100 μg/ml ampicillin at 37°C with shaking until the culture reached an optical density at 600 nm (OD600) of 0.5, at which point 0.2 μg/mL anhydrotetracycline (aTc) were added to induce *estDZ4* overexpression. The cells were collected, lysed by brief sonication and the clarified lysates were used to run a zymogram as described before (Zarafeta et al., 2016b). Briefly, the gel was rinsed in distilled water and incubated for 30 min at 37°C in 0.1% Fast Red TR-salt (4-chloro-2-methylbenzenediazonium salt) in 0.1 M Tris–HCl buffer (pH 7.0) containing 2% of a 1% (v/v) 1-naphthyl acetate solution in acetone. Esterolytic activity was visualized by the appearance of a band of red-brown color.

For EstDZ4 purification, Origami 2 (DE3) cells carrying the pASK-EstDZ4 plasmid were grown as described above, and the induction of *estDZ4* overexpression was performed by overnight incubation at 25°C with shaking. The cells from a 500 mL culture grown in a 2 L shake flask were harvested, washed, re-suspended in 10 mL equilibration buffer enhanced with 1% Triton X-100 (v/v) and lysed by brief sonication steps on ice. The cell extract was clarified by centrifugation at 10,000 × *g* for 15 min at 4°C and the supernatant was collected and mixed with 0.5 mL Ni-NTA agarose beads (Qiagen – Hilden, Germany) and agitated mildly for 1 h at 4°C. The mixture was then loaded onto a 5 mL polypropylene column (Thermo Fisher Scientific – Waltham, United States), the flow-through was discarded, and the column was washed with 10 mL of NPI20/ Triton wash buffer followed by a second wash with non-Triton-enriched NPI20. EstDZ4 was eluted using 1 mL of NPI200 elution buffer. All buffers used for purification were prepared according to the manufacturer's protocol (Qiagen – Hilden, Germany) unless stated otherwise. EstDZ4 was further purified by size-exclusion chromatography (SEC) using a Superdex75 10/300GL column (GE Healthcare, United States).

Protein concentration was estimated according to the assay described by Bradford (Bradford, 1976) using bovine serum albumin as a standard. The purified protein was visualized by SDS-PAGE analysis and staining with Coomassie blue or western blotting using an anti-polyhistidine monoclonal antibody conjugated with horseradish peroxidase (Sigma – St. Louis, MO, United States).

The catalytic activity of EstDZ4 was determined by quantification of the amount of *p*-nitrophenol (pNP) released from pNP ester substrates by photometric measurement at 410 nm. Unless stated otherwise, the 100 μL standard reaction mixture consisted of 25 mM phosphate buffer pH 6.5 enriched with 0.05% Triton X-100 (v/v), 2 mM pNP-octanoate and 1 μg/mL of pure enzyme, and was carried out for 5 min at 75°C on a MJ Research thermal cycler, with a pre-incubation setting of the buffer to the target temperature, before the enzyme was added. Enzymic activity was recorded using a Safire II-Basic plate reader (Tecan, Austria) by measuring the absorbance of the released pNP at 410 nm, immediately after the reaction was completed. For the substrate specificity experiments, a range of different pNP-fatty acyl esters, such as acetate (C2), butyrate (C4), octanoate (C8), decanoate (C10), laurate (C12), and palmitate (C16) were used in the standard reaction. For the determination of the enzyme's optimal pH, reactions were carried out at 40°C in 25 mM acetate, phosphate, Tris–HCl and glycine-NaOH buffers for pH values 4–6, 6–7, 7–9, and 9–10, respectively. Activity was measured by recording absorbance at 348 nm, the isosbestic point of pNP, so as to exclude the pH effect on the readings. Temperature profiling of EstDZ4 was performed by incubating the standard reaction at temperatures ranging from 25 to 70°C. Thermostability experiments were conducted by incubating the enzyme at high temperatures for prolonged time periods and subsequently measuring its activity in the standard reaction.

RESULTS AND DISCUSSION

Development of ANASTASIA and Stand-Alone Portability

Automated nucleotide aminoacid sequences translational plAtform for systemic interpretation and analysis was developed as a user-friendly, web-based, computational pipeline, aiming to address numerous and diversified series of metagenomic analysis tasks for the massive characterization of a broad constellation of bacterial metagenomes. For this scope it integrates numerous bioinformatic tools (**Figures 1**, **2**), as components of automated workflows (**Figure 3**) designed to handle a wide range of processing tasks. In contrast to the limited versatility and automation capacity of other metagenomic pipelines, ANASTASIA's automated workflows are tackling various processing steps, from the handling of raw sequencing data to the putative function predictions for gene encoding sequences and the powerful functional characterization of the underlying emerging molecular networks (**Figure 4**). ANASTASIA can address different scenarios, from the screening of thermophiles to the systematic screening of the human microbiome in various infections, as part of the Hellenic Bioinformatic computational Infrastructure ELIXIR-GR, which represents the Hellenic node of ELIXIR. Below, as a showcase of

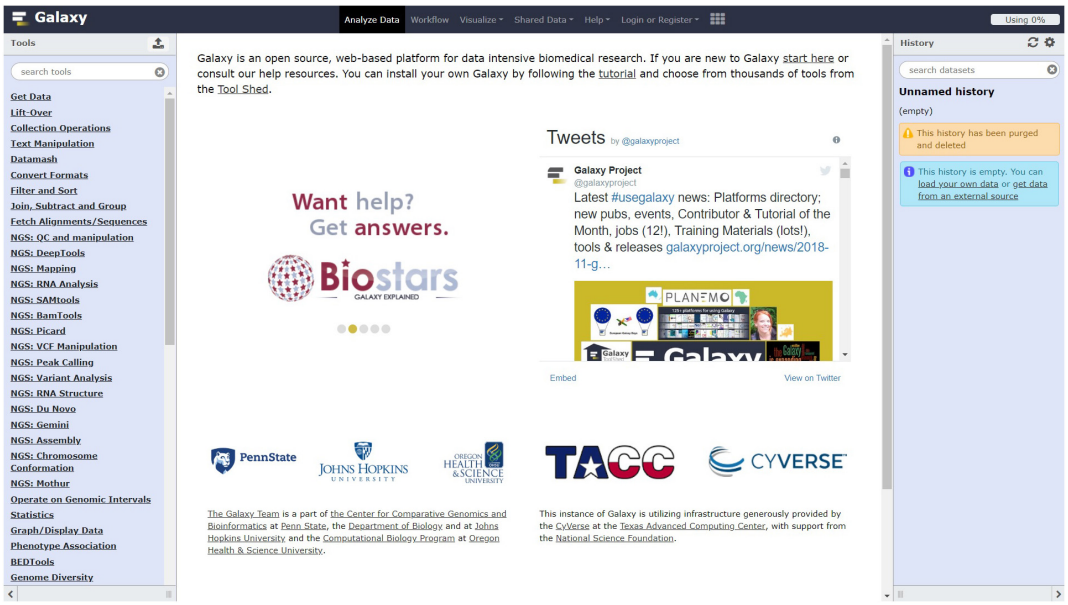


FIGURE 1 | The original Galaxy front end.

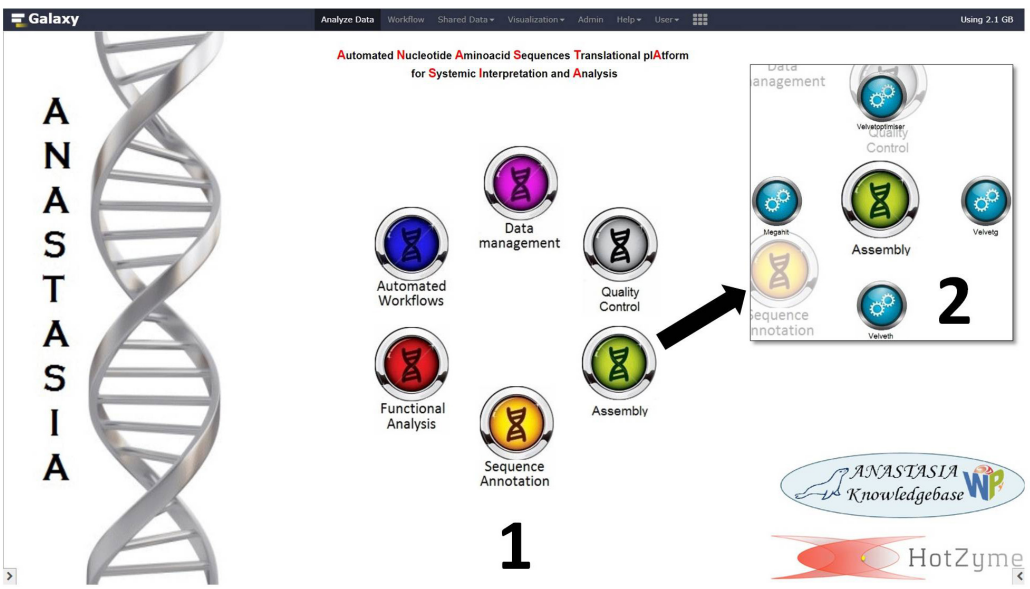


FIGURE 2 | The customized front page of ANASTASIA framework with (1) the main menu buttons and (2) the submenu of "Assembly" suite of tools. The submenu consists of 4 tools that handle NGS assembly from reads to contigs and a tool that imports the output assembly data into a MariaDB database.

the practical utility of ANASTASIA as an efficient platform for data-driven discovery, we report the successful *in silico* mining of a novel thermostable esterase, termed EstDZ4, as a result of the exhaustive analysis of a metagenomic sample taken from a hot spring located in Krisuvik, Iceland.

Access to ANASTASIA and its tools is possible either via the publicly available server[4] or as a local bundle installation after download by executing a Bash script[5]. The script, which works in any Linux platform, performs the following actions: (i) automatically downloads the most recent (without compatibility issues) Galaxy instance; (ii) assigns a MariaDB database for this instance; (iii) builds the ANASTASIA interface and configures the web server for online viewing of the platform; (iv) downloads the customized tools and integrates them in

[4]http://motherbox.chemeng.ntua.gr/anastasia_dev/

[5]https://bitbucket.org/TYRANISTAR/anastasia/

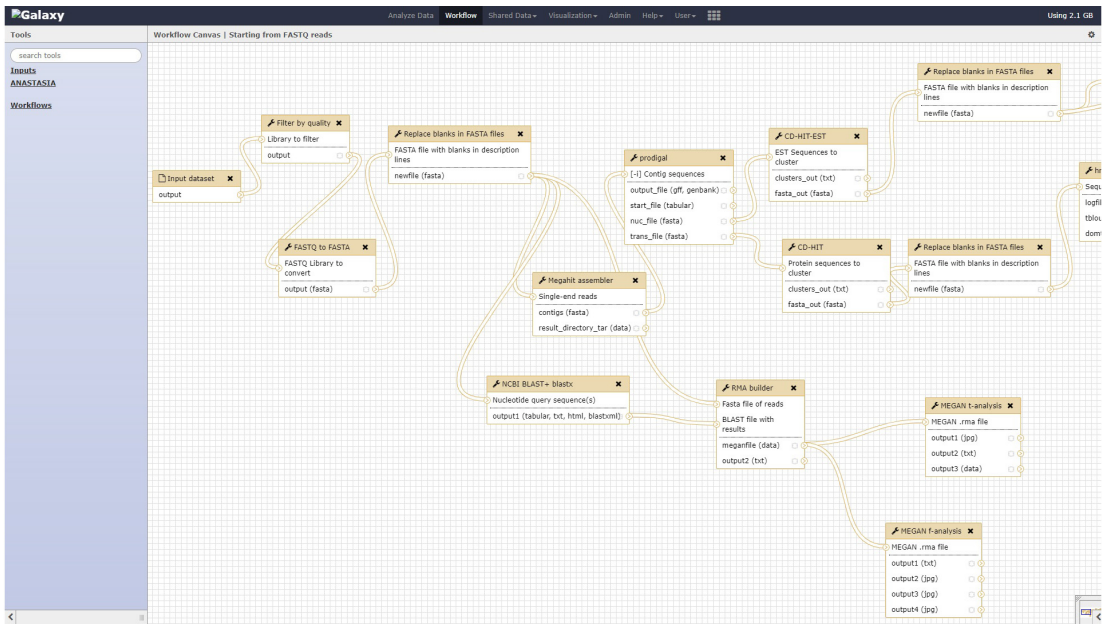


FIGURE 3 | Galaxy's Workflow Canvas.

ANASTASIA; (v) downloads any essential databases needed for the tools (NCBI-nr, nt, UniProt, etc.) to operate and formats them accordingly to be usable (e.g., makeblastdb tool for creating BLAST-able databases) and (vi) runs the Galaxy startup script (sh run.sh) in order for ANASTASIA to activate.

Discovery and Biochemical Characterization of EstDZ4

Automated nucleotide aminoacid sequences translational plAtform for systemic interpretation and analysis was used to analyze the metagenomic DNA of the Is3-13 sample originating from a high temperature pool (90°C, pH 3.5–4.0) in Krisuvik, Iceland (Menzel et al., 2015) described above. From this analysis, a specific sequence was selected for further investigation as a proof of concept for the ability of the platform to identify new enzymes. The selected sequence, named *EstDZ4*, was predicted to encode a 454-amino acid protein with a predicted molecular mass of 46.3 kDa. According to a BLAST analysis against the SwissProt/Uniprot database containing characterized proteins, EstDZ4 presented 23% identity (query coverage 84%) to a previously characterized GDSL esterase/lipase from *A. thaliana* [UniProtKB/Swiss-Prot: Q9FIA1.1] (Cheng et al., 2017) and a 99% identity (query coverage 98%) to a putative lipase/esterase from *Thiomonas sp.CB3* [GenBank: CQR41430.1]. The same analysis assigned the protein to the Triacylglycerol lipase-like subfamily of the SGNH hydrolases [NCBI Conserved Domains Datatabase accession number: cl01053 (Marchler-Bauer et al., 2017)], which is a diverse family of lipases and esterases. Sequence analysis against the Pfam-A database using HMMER predicted that the sequence contains a GDSL-like Lipase/Acylhydrolase catalytic domain spanning amino acids 61-442, as well as a signal peptide (amino acids 1-38). Further examination with EFICAz assigned a putative esterolytic activity to the sequence, according to its EC number prediction (3.1.1.).

The *estDZ4* gene was amplified by PCR from the Is3-13metagenomic DNA sample and was cloned (after excluding the predicted signal peptide) into the plasmid pASK75 (Skerra, 1994) to form the vector pASK-EstDZ4. *E. coli* Origami 2 (DE3) cells were transformed with pASK-EstDZ4 and the recombinant protein was produced as described in the Section "Materials and Methods." To examine the putative esterolytic activity of EstDZ4, a zymogram analysis was performed using clarified lysates of cells producing the recombinant protein and cells carrying an empty vector. Following the separation of the proteins contained in the clarified cell lysates by native PAGE, the gel was exposed to 1-naphtyl acetate as a potential substrate for ester hydrolysis and stained with Fast Red. Immediately upon staining, a band of red-brown color appeared only for the EstDZ4-producing sample (**Figure 5A**), thus indicating that EstDZ4 is an esterolytic enzyme.

EstDZ4 was then purified in soluble form by immobilized metal affinity chromatography (IMAC) followed by SEC (data not shown) and pure protein was used for all subsequent biochemical characterization experiments. Substrate specificity of the new esterase was evaluated by performing enzymic reactions using pNP esters of fatty acids of various lengths as substrates in the standard reaction. As shown in **Figure 5B**, EstDZ4 exhibits esterolytic activity against small and medium chain lengths (C2–C12), with an apparent preference for pNP-octanoate.

Assaying the esterolytic activity of EstDZ4 within the pH range of 4–10 at 40°C using pNP-octanoate as the substrate, revealed that the optimal pH for the new enzyme is pH 6.5

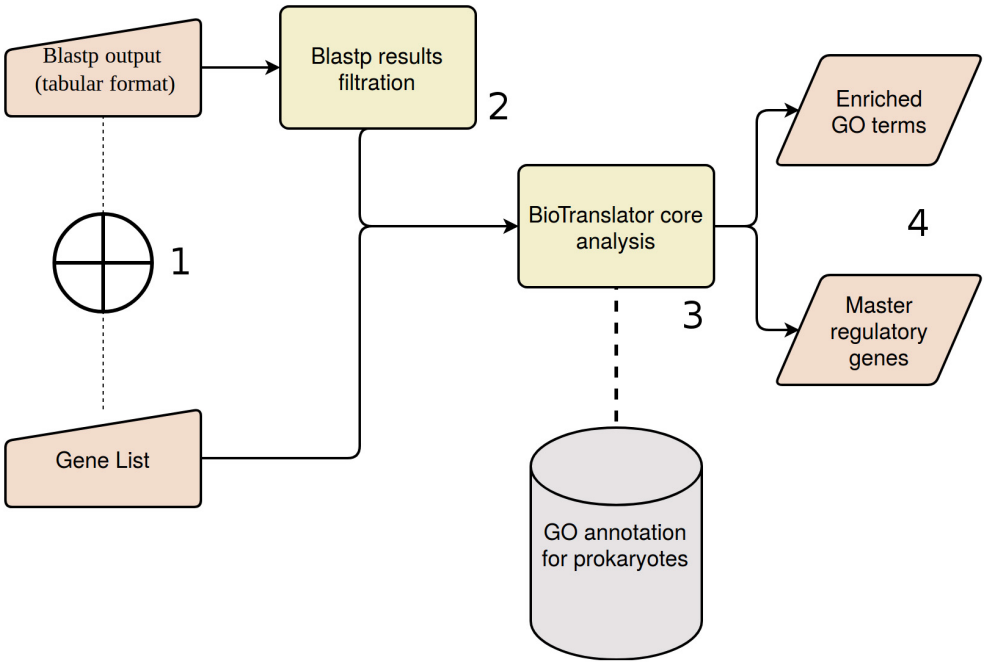


FIGURE 4 | BioTranslator workflow. (1) User could start the analysis with a pre-defined gene list, or sequence similarities derived from Blastp analysis. (2) Blastp results are filtered if the initial input is the respective tabular file. (3) Pathway analysis and master regulators detection are based on BioTranslator core algorithms. (4) Final outputs of BioTranslator analysis.

(**Figure 6A**). Measurements of its relative catalytic activity at different temperatures, on the other hand, showed that EstDZ4 has a broad temperature range of action as it retains high levels of esterolytic activity at temperatures between 40 and 85°C, with an optimal temperature of action at 75°C (**Figure 6B**). In order to evaluate the thermostability of EstDZ4, the enzyme was incubated for prolonged time periods in high temperatures and its residual activity was measured. As shown in **Figure 6C**, EstDZ4 exhibited a half-life of ∼5 h when exposed to 80°C, and even after 24 h of incubation at 70 and 75°C, the enzyme retained more than 40% of its initial activity, demonstrating that EstDZ4 is a highly thermostable esterase.

Novelty of EstDZ4

The characterization of EstDZ4 as a novel enzyme does not imply the discovery of a new biological catalytic reaction or a new microbial species (although it might very well have originated from one) but derives from the fact that an enzyme that had not been described before (only 23% similar to its closest characterized entry in UniProt/Swissprot database) was isolated from a vast gene pool of numerous different microbial species originating from the same environmental sample. Bioinformatic-based predictions concerning enzymatic function are based either on the entirety of the sequence or on smaller conserved motifs (e.g., protein domains) of already known enzymes so the discovery of a totally novel enzymatic activity is a rather unachievable task if tackled solely by *in silico* approaches. However, these approaches remain the cornerstone of such an endeavor as their lists of potential enzymes provide the starting point for further curation via wet lab protocols. Such a curation aims not only to confirm each gene product's predicted enzymatic function and to pinpoint its optimum conditions (temperature, pH, etc.) but also to search for new additional putative target substrates, potentially uncovering novel enzymatic activities. Similarly, a novel species cannot be determined with 100% certainty only by the bioinformatic paradigm, although several strategies have emerged (Droge and McHardy, 2012). In order to discover and properly characterize a new species, prior culturing is needed but this is not usually possible for most of the microbial species in an environmental niche. Nevertheless, this is exactly the type of issue that ANASTASIA's pipelines aim to bypass. Detecting novel gene products, such as EstDZ4, from metagenomic samples whose species cannot be cultured and assigning functional properties to these sequences is one of the many different objectives of this platform. The choice of metagenomic sample to mine for enzymes of potential interest is equally important. One of the key features of EstDZ4 that makes it highly interesting is its thermostability which can be attributed to the physical properties of the site where the metagenomic material was sampled from.

Comparison of ANASTASIA With Other Metagenomic Solutions

The idea of building automated workflows for metagenomic analysis is not a new one and there is a number of older (Ladoukakis et al., 2014) and newer solutions available (Kultima et al., 2016; Lugli et al., 2016). Most of these solutions, however, represent mostly academic compilations, requesting from their

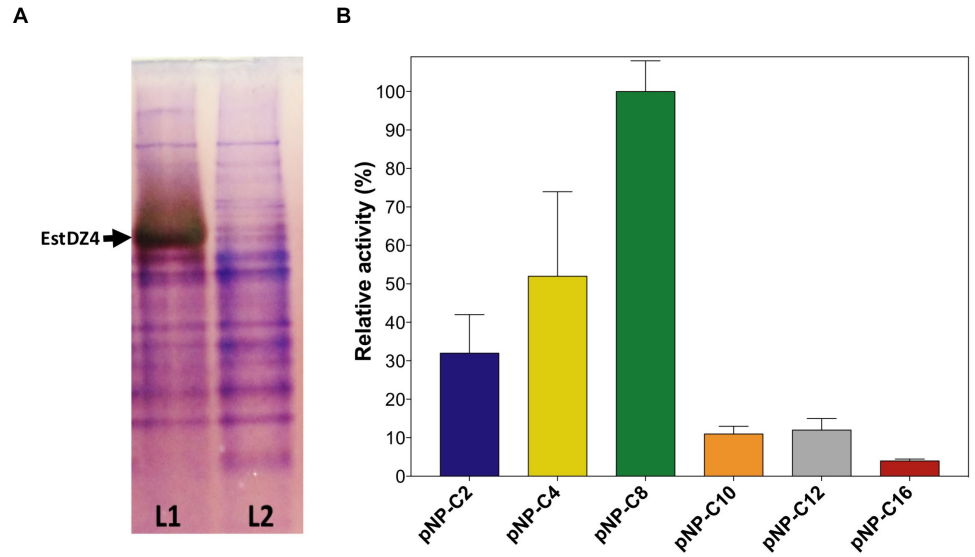


FIGURE 5 | Confirmation of EstDZ4 esterolytic activity. **(A)** Detection of EstDZ4 esterolytic activity via native PAGE analysis and Fast Red staining using 1-naphthyl acetate as a substrate. L1, clarified lysate of cells overexpressing estDZ4; L2, clarified lysate of cells carrying an empty vector. **(B)** Substrate profiling of the esterolytic activity of EstDZ4, using purified enzyme. The relative enzymic activity was measured as released pNP at 410 nm (pH 7, 40°C). The reported values correspond to the mean value of three independent experiments performed in triplicate and the error bars to one standard deviation from the mean value. Assaying the esterolytic activity of EstDZ4 within the pH range of 4–10 at 40°C using pNP-octanoate as the substrate, revealed that the optimal pH for the new enzyme is pH 6.5 (**Figure 6A**). Measurements of its relative catalytic activity at different temperatures, on the other hand, showed that EstDZ4 has a broad temperature range of action as it retains high levels of esterolytic activity at temperatures between 40–85°C, with its optimal temperature of action being at 75°C (**Figure 6B**). In order to evaluate the thermostability of EstDZ4, the enzyme was incubated for prolonged time periods in high temperatures and its residual activity was measured. As shown in **Figure 6C**, EstDZ4 exhibited a half-life of ∼5 h when exposed to 80°C, and even after 24 h of incubation at 70 and 75°C, the enzyme retained more than 40% of its initial activity, demonstrating that EstDZ4 is a highly thermostable esterase.

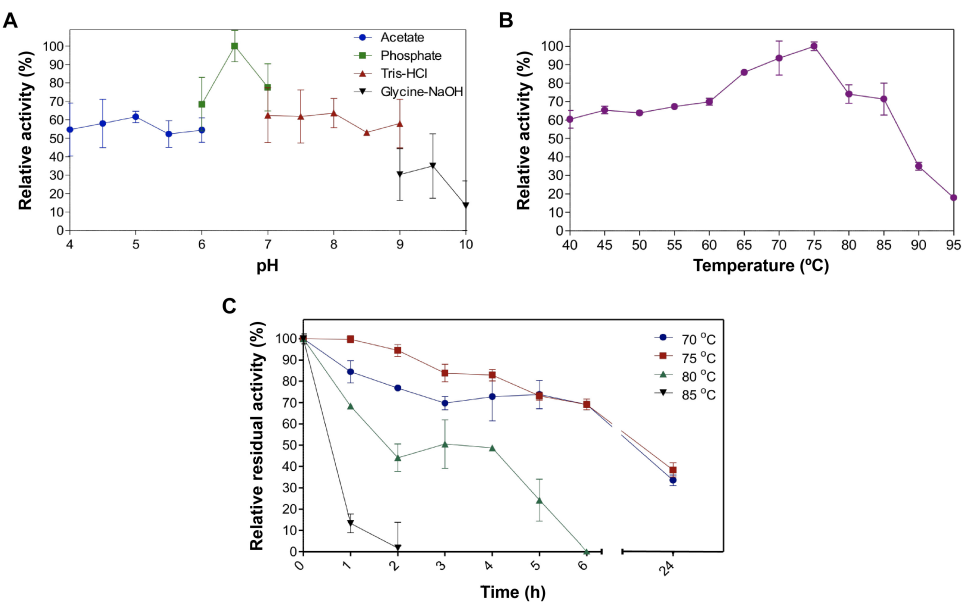


FIGURE 6 | Biochemical properties of EstDZ4. **(A)** Effect of pH on EstDZ4 activity. Enzymic activity was measured in the standard reaction, at pH values ranging from 4 to 10, using the indicated buffers. **(B)** Effect of temperature on EstDZ4 activity. Enzymic activity was measured at temperatures ranging from 40 to 95°C in the standard reaction. **(C)** Thermostability of EstDZ4 was evaluated by measurements of its esterolytic activity in the standard reaction after exposure to 75, 75, 80, and 85°C for up to 24 h. The reported values correspond to the mean value from three independent experiments performed in triplicate and the error bars to one standard deviation from the mean value.

users a varying degree of familiarity with (bio)informatics programming/scripting, in order to be able to install, execute, and properly parse their resulting datasets since they operate only via Linux command line (Li, 2009; Treangen et al., 2013; Kultima et al., 2016), and some of them even comprise tools or workflows that focus only on specific parts of the metagenomic analysis (see **Table 1**). A very recent example is another galaxy-based platform: ASAIM (Batut et al., 2018), which includes integrated tools for every part of a complete metagenomic analysis from sequence assembly to sequence annotation but its automated workflows don't include most of them as they mainly focus on taxonomic and functional analysis, leaving the task of novel enzyme mining up to the user's experience in dealing with the rest of stand-alone tools. Even the Galaxy platform, which ANASTASIA has based its development on, has a limited arsenal of integrated tools for metagenomic analysis on its official public server[6] and it requires an experienced user to download and install the platform in order to customize it accordingly for a complete analysis. On the other hand ANASTASIA offers an all-inclusive set of tools and automated workflows (**Figure 7**) tuned to be used to tackle each separate task of a metagenomic study, from assembling raw sequencing reads to fully annotating and predicting the function of the gene coding sequences derived from each sample. In addition our platform can be easily used even by the most inexperienced researcher, as it is available on a public server (motherbox.chemeng.ntua.gr) via an intuitive graphic user interface and offers built-in ready-to-use automated workflows with default parameters that

[6]https://usegalaxy.org/

TABLE 1 | Comparison of current bioinformatic pipelines for metagenomic data analysis with ANASTASIA.

Tasks \ Pipeline	Quality control	Assembly	Gene detection	Functional annotation	Taxonomic analysis	Data management system	Graphical user interface	Customizable
CloVR-metagenomics	x	x	✓	✓	✓	✓	✓	x
Galaxy platform	✓	x	x	x	✓	x	✓	x
IMG/M	x	x	✓	✓	✓	✓	✓	x
MetAMOS	✓	✓	✓	✓	✓	x	x	x
MG-RAST	✓	x	✓	✓	✓	✓	✓	x
RAMMCAP	x	x	✓	✓	✓	x	x	x
SmashCommunity	✓	✓	✓	✓	✓	✓	x	x
MOCAT2	✓	✓	✓	✓	✓	x	x	x
MEGAnnotator	✓	✓	✓	✓	x	x	x	x
ASAIM	✓	✓	✓	✓	✓	x	✓	✓
ANASTASIA	✓	✓	✓	✓	✓	✓	✓	✓

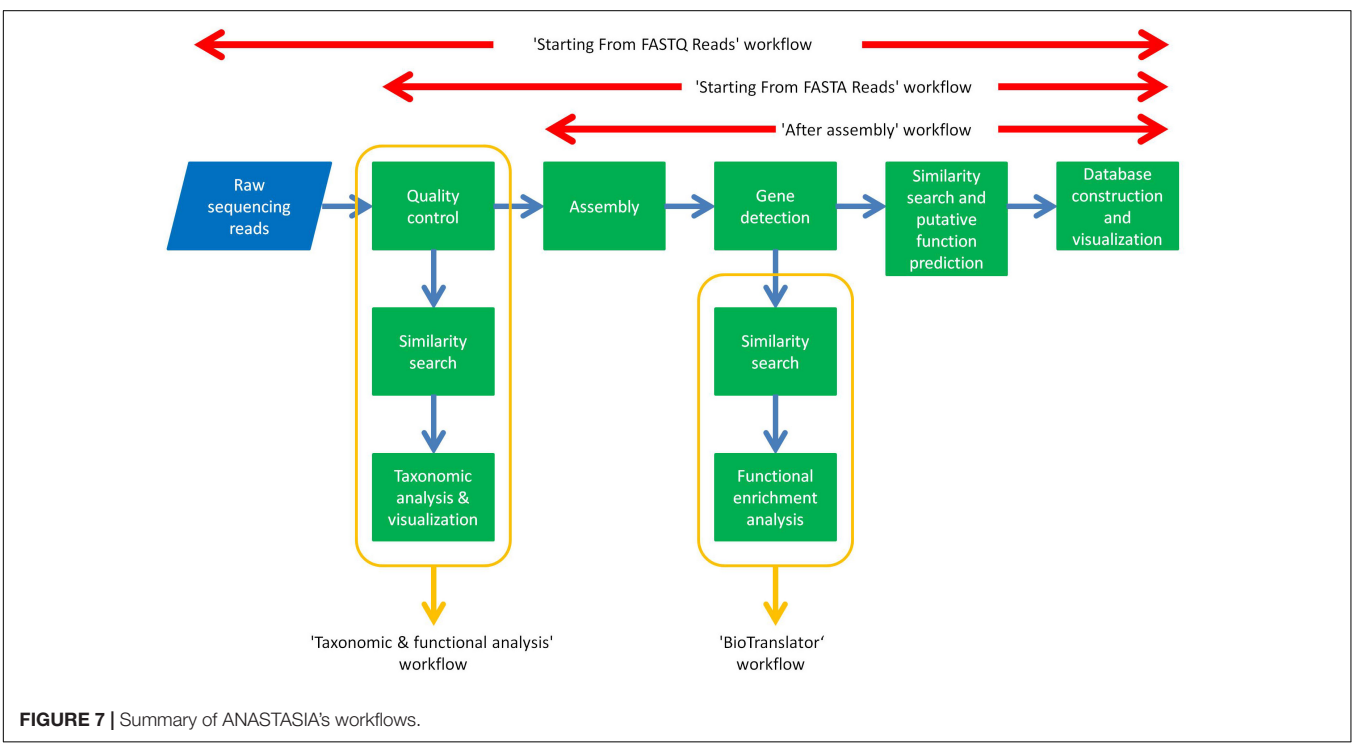


FIGURE 7 | Summary of ANASTASIA's workflows.

can handle most type of datasets. This setup also includes the online data management system mentioned before, that allows the user to manually access and curate the otherwise hard-to-handle resulting datasets from an analysis. In accordance with most command-line pipelines and its Galaxy predecessor, ANASTASIA can also be locally downloaded, installed in any Linux system and even be further customized by highly proficient, bioinformatic, power users providing the same graphic user interface and data management system as on the public server. **Table 1** demonstrates the strengths of the platform in comparison with the most common open-source metagenomic pipelines already available.

CONCLUSION

Sequencing technologies have emerged to become an indispensable tool for metagenomics revolutionizing the ways, through which we can probe an environmental niche and extract more inclusive information about its genomic content for an ever decreasing cost. This evolution however, is followed by the immense increase of the generated amount of data and the need for numerous different bioinformatic tasks, which are essential for its complete annotation. Here, we present a powerful solution to these issues, the ANASTASIA platform. ANASTASIA is a portable web repository for large metagenomic datasets, providing automatic, bioinformatic workflows for data handling, and major annotation tasks via a friendly GUI. ANASTASIA functions as an intuitive and easily accessible tool, both for biologists and other users needing to store, manage and fully annotate very large metagenomic datasets, as those generated and compiled during the HotZyme project, where it was utilized for the exhaustive analysis of sequencing data from various metagenomic samplings around the world. ANASTASIA, as an automated analytical platform, represents a stable and well tested environment for the future integration of families of newer and faster algorithms, addressing diverse bioinformatic tasks emerging as pressing needs in current, ever-increasing in complexity, and metagenomic studies.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The work was co-funded by the EU/FP7/KBBE-2010.3.5-04 Microbial diversity and metagenomic mining for biotechnological innovation, "Systematic screening for novel hydrolases from hot environments" project (HotZyme – Grant Agreement: 265933), and the e-NIOS Applications P.C. TK was funded by the project "ELIXIR-GR: The Greek Research Infrastructure for Data Management and Analysis in Life Sciences" (MIS 5002780). ELIXIR was partly funded by the European Commission within the Research Infrastructures program of Horizon 2020. DZ was the recipient of a Ph.D. fellowship from the Greek State Scholarships Foundation (Idryma Kratikon Ypotrofion-IKY) in the framework of the Research Grant Excellence IKY-Siemens Program, which was co-financed by the European Social Fund and the Greek Government. EP was the recipient of a post-doctoral research fellowship from the State Scholarships Foundation (IKY), funded by the Action "Support for post-doctoral researchers," funded from the OP "Human Resources Development, Education and Lifelong Learning," with priority axes 6,8,9, co-funded by the European Social Fund (ESF) and the Greek State. The estDZ4 nucleotide sequence has been deposited in GenBank under the accession code MK942067.

ACKNOWLEDGMENTS

We thank Prof. Anders Krogh, from University of Copenhagen for providing access to their university server in order to conduct a series of tests for our numerous bioinformatic tools and platforms and install our beta version of ANASTASIA. Special thanks to Georgios Papadopoulos for his help and feedback on designing the interactive front-end of ANASTASIA. We also thank Frantzeska Siettou for technical assistance.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1006/jmbi.1990.9999

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29.

Batut, B., Gravouil, K., Defois, C., Hiltemann, S., Brugere, J. F., Peyretaillade, E., et al. (2018). ASaiM: a galaxy-based framework to analyze microbiota data. *Gigascience* 7, doi: 10.1093/gigascience/giy057

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1006/abio.1976.9999

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2:10. doi: 10.1186/2047-217X-2-10

Chatziioannou, A. A., and Moulos, P. (2011). Exploiting statistical methodologies and controlled vocabularies for prioritized functional analysis of genomic experiments: the StRAnGER web application. *Front. Neurosci.* 5:8. doi: 10.3389/fnins.2011.00008

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* 89, 789–804. doi: 10.1111/tpj.13415

Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (1998). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 63, 41–49. doi: 10.1016/j.ymeth.2013.06.027

Droge, J., and McHardy, A. C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.* 13, 646–655. doi: 10.1093/bib/bbs031

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* 21, 2224–2241. doi: 10.1101/gr.126599.111

Escobar-Zepeda, A., Lozano-Godoy, E. E., Raggi, L., Segovia, L., Merino, E., Gutierrez, R. M., et al. (2018). Analysis of sequencing strategies and tools for taxonomic annotation: defining standards for progressive metagenomics. *Sci. Rep.* 8:12034. doi: 10.1038/s41598-018-30515-5

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185. doi: 10.1101/gr.8.3.175

Extensible Markup Language (2013). Available from: http://www.w3.org/XML/ (accessed April 5, 2017).

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

FastQC (2010). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed April 5, 2017).

Fastx Toolkit (2009). Available from: http://hannonlab.cshl.edu/fastx_toolkit/ (accessed April 5, 2017).

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Galaxy Tool Shed (2015). Available from: https://toolshed.g2.bx.psu.edu/ (accessed April 5, 2017).

Garrity, G. M. (2016). A new genomics-driven taxonomy of bacteria and archaea: are we there yet? *J. Clin. Microbiol.* 54, 1956–1963. doi: 10.1128/JCM.00200-16

Gerlt, J. A., and Babbitt, P. C. (2000). Can sequence determine function? *Genome Biol.* 1:REVIEWS0005.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86. doi: 10.1186/gb-2010-11-8-r86

Gross, A., Pruski, C., and Rahm, E. (2016). Evolution of biomedical ontologies and mappings: overview of recent approaches. *Comput. Struct. Biotechnol. J.* 14, 333–340. doi: 10.1016/j.csbj.2016.08.002

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Javascript (2016). Available from: https://www.javascript.com/ (accessed January 17, 2017).

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Kultima, J. R., Coelho, L. P., Forslund, K., Huerta-Cepas, J., Li, S. S., Driessen, M., et al. (2016). MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32, 2520–2523. doi: 10.1093/bioinformatics/btw183

Kumar, N., and Skolnick, J. (2012). EFICAz(2.5): application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics* 28, 2687–2688. doi: 10.1093/bioinformatics/bts510

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578. doi: 10.1128/MMBR.00009-08

Ladoukakis, E., Kolisis, F. N., and Chatziioannou, A. A. (2014). Integrative workflows for metagenomic analysis. *Front. Cell Dev. Biol.* 2:70. doi: 10.3389/fcell.2014.00070

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10:359. doi: 10.1186/1471-2105-10-359

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genomics* 11, 25–37. doi: 10.1093/bfgp/elr035

Lugli, G. A., Milani, C., Mancabelli, L., van Sinderen, D., and Ventura, M. (2016). MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol. Lett.* 363:fnw049. doi: 10.1093/femsle/fnw049

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi: 10.1093/nar/gkw1129

Menzel, P., Gudbergsdottir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., et al. (2015). Comparative metagenomics of eight geographically remote terrestrial hot springs. *Microb. Ecol.* 70, 411–424. doi: 10.1007/s00248-015-0576-9

Mercurial SCM (2005). Available at: http://mercurial.selenic.com/ (accessed April 5, 2017).

Moutselos, K., Maglogiannis, I., and Chatziioannou, A. (2011). GOrevenge: a novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies. *IEEE Trans. Biomed. Eng.* 58, 3522–3527. doi: 10.1109/TBME.2011.2164794

MySQL (1995). Available at: https://www.mysql.com/ (accessed June 12, 2015).

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi: 10.1093/dnares/dsn027

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi: 10.1093/nar/gki866

Pandey, R. V., Pabinger, S., Kriegner, A., and Weinhausel, A. (2016). ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. *BMC Bioinformatics* 17:56. doi: 10.1186/s12859-016-0915-y

Pierro, M. D. (2010). web2py for scientific applications. *Comput. Sci. Eng.* 13, 64–69. doi: 10.1109/mcse.2010.97

Rho, M., Tang, H., and Ye, Y. (2010). (FragGeneScan): predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/s0168-9525(00)02024-2

Scholz, M. B., Lo, C. C., and Chain, P. S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15. doi: 10.1016/j.copbio.2011.11.013

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Skerra, A. (1994). Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in *Escherichia coli*. *Gene* 151, 131–135. doi: 10.1016/0378-1119(94)90643-2

Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins Struct. Funct. Genet.* 28, 405–420. doi: 10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* 11:207. doi: 10.1186/gb-2010-11-5-207

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

The Apache (1995). *Software Foundation*. Available from: http://httpd.apache.org/ (accessed March 20, 2017).

The MariaDB Foundation (2009). Available from: https://mariadb.org/ (accessed June 12, 2015).

The Perl Programming Language (2002). Available from: http://www.perl.org (accessed November 6, 2016).

The Python Programming Language (2001). Available from: http://www.python.org (accessed October 30, 2018).

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

van der Walt, A. J., van Goethem, M. W., Ramond, J. B., Makhalanyane, T. P., Reva, O., and Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics.* 18:521. doi: 10.1186/s12864-017-3918-9

Veloso, F., Riadi, G., Aliaga, D., Lieph, R., and Holmes, D. S. (2005). Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. *OMICS* 9, 91–105. doi: 10.1089/omi.2005.9.91

Vollmers, J., Wiegand, S., and Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS One* 12:e0169662. doi: 10.1371/journal.pone.0169662

Wilkening, J., Wilke, A., Desai, N., and Meyer, F. (2009). "Using Clouds for Metagenomics: A Case Study," in *IEEE International Conference on Cluster Computing and Workshops*, (New Orleans, LA: IEEE).

Wohlgemuth, R., Littlechild, J., Monti, D., Schnorr, K., von Rossum, T., Siebers, B., et al. (2018). Discovering novel hydrolases from hot environments. *Biotechnol. Adv.* 36, 2077–2100. doi: 10.1016/j.biotechadv.2018.09.004

Zarafeta, D., Kissas, D., Sayer, C., Gudbergsdottir, S. R., Ladoukakis, E., Isupov, M. N., et al. (2016a). Discovery and characterization of a thermostable and highly halotolerant GH5 Cellulase from an icelandic hot spring isolate. *PLoS One* 11:e0146454. doi: 10.1371/journal.pone.0146454

Zarafeta, D., Moschidi, D., Ladoukakis, E., Gavrilov, S., Chrysina, E. D., Chatziioannou, A., et al. (2016b). Metagenomic mining for thermostable esterolytic enzymes uncovers a new family of bacterial esterases. *Sci. Rep.* 6:38886. doi: 10.1038/srep38886

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275

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.

Copyright © 2019 Koutsandreas, Ladoukakis, Pilalis, Zarafeta, Kolisis, Skretas and Chatziioannou. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 19 March 2019
doi: 10.3389/fgene.2019.00159

A Multi-Cohort and Multi-Omics Meta-Analysis Framework to Identify Network-Based Gene Signatures

Adib Shafi[1], *Tin Nguyen*[2], *Azam Peyvandipour*[1], *Hung Nguyen*[2] *and Sorin Draghici*[1,3]*

[1] Department of Computer Science, Wayne State University, Detroit, MI, United States, [2] Department of Computer Science and Engineering, University of Nevada, Reno, NV, United States, [3] Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, United States

Although massive amounts of condition-specific molecular profiles are being accumulated in public repositories every day, meaningful interpretation of these data remains a major challenge. In an effort to identify the biomarkers that describe the key biological phenomena for a given condition, several approaches have been developed over the past few years. However, the majority of these approaches either (i) do not consider the known intermolecular interactions, or (ii) do not integrate molecular data of multiple types (e.g., genomics, transcriptomics, proteomics, epigenomics, etc.), and thus potentially fail to capture the true biological changes responsible for complex diseases (e.g., cancer). In addition, these approaches often ignore the heterogeneity and study bias present in independent molecular cohorts. In this manuscript, we propose a novel multi-cohort and multi-omics meta-analysis framework that overcomes all three limitations mentioned above in order to identify robust molecular subnetworks that capture the key dynamic nature of a given biological condition. Our framework integrates multiple independent gene expression studies, unmatched DNA methylation studies, and protein-protein interactions to identify methylation-driven subnetworks. We demonstrate the proposed framework by constructing subnetworks related to two complex diseases: glioblastoma and low-grade gliomas. We validate the identified subnetworks by showing their ability to predict patients' clinical outcome on multiple independent validation cohorts.

Keywords: multi-cohort, multi-omics, meta-analysis, subnetwork identification, GBM, LGG

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council (CNR), Italy

Reviewed by:
Alfredo Pulvirenti,
Università degli Studi di Catania, Italy
Hamed Bostan,
North Carolina State University, United States

***Correspondence:**
Sorin Draghici
sorin@wayne.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *02 November 2018*
Accepted: *14 February 2019*
Published: *19 March 2019*

Citation:
Shafi A, Nguyen T, Peyvandipour A, Nguyen H and Draghici S (2019) A Multi-Cohort and Multi-Omics Meta-Analysis Framework to Identify Network-Based Gene Signatures. Front. Genet. 10:159. doi: 10.3389/fgene.2019.00159

1. INTRODUCTION

Due to the rapid advances in high-throughput technologies, massive amounts of biological data are currently available in public repositories for many diseases. These biological data include various *omics* profiles such as genomic, transcriptomic, metabolomic, and proteomic data, each of which describes different aspects of cellular mechanisms. Understanding the *mechanism of action* for a given disease from these vast resources and subsequently identifying reliable biomarkers that can predict the patients' clinical outcome has become a major challenge.

Over the last decade, the number of disease-specific biomarkers reported by different research groups has increased exponentially. However, biomarkers obtained from different studies of the same condition often show very poor agreement with each other (Ein-Dor et al., 2006). As a result, only a few of the proposed biomarkers are currently in clinical use (Burke, 2016). One of the primary reasons for this *reproducibility crisis* is that many of the conventional biomarker discovery

methods simply rely on selecting a list of candidate genes based on their differential expression across the given phenotypes (disease vs. normal, treated vs. non-treated, subtype A vs. subtype B, etc). Better results can be obtained by utilizing gene interaction data that became available with the introduction of publicly available sources such as pathway knowledge databases [e.g., KEGG (Ogata et al., 1999; Kanehisa and Goto, 2000), Reactome (Matthews et al., 2009)] or protein-protein interaction databases [e.g., HPRD (Peri et al., 2003), STRING (Szklarczyk et al., 2016)].

Numerous computational methods have been proposed that aim to address the above-mentioned challenge by integrating known interactions between the genes and subsequently identifying network-based markers using different strategies. For instance, PinnacleZ (Chuang et al., 2007) and DIAMOnD (Ghiassian et al., 2015) use greedy algorithm-based techniques; jActiveModules (Ideker et al., 2002) and COSINE (Ma et al., 2011) utilize evolutionary algorithms; HotNet (Vandin et al., 2011) and ResponseNet (Lan et al., 2011) use diffusion-flow based techniques; EnrichNet (Glaab et al., 2012) employs random walk algorithms; etc. These network-based approaches have been reviewed elsewhere (Mitra et al., 2013; Nguyen T. et al., 2018). It has been demonstrated in various disease conditions [e.g., breast cancer (Chuang et al., 2007), colorectal cancer (Shi et al., 2012; Shafi et al., 2015), and ovarian cancer (Jin et al., 2015)] that network-based markers are more reproducible and reliable for predicting patients' clinical outcome than individual gene biomarkers. Although somewhat useful, the majority of these methods construct their networks using only one transcriptomic experiment. Therefore, they are unable to account for the heterogeneity that may arise due to the biological and technical variabilities present in independent studies of a given disease (Drăghici et al., 2006; MAQC Consortium, 2006).

In order to account for the data heterogeneity present in the individual studies, several meta-analysis approaches have been proposed over the past years. These can be divided into two main categories. The approaches in the first category use multiple sample-unmatched studies of the same data type (e.g., mRNA) and aim to identify robust gene signatures that can distinguish disease-affected individuals from the healthy ones. These approaches include classical *p*-value-based approaches (Fisher, 1925; Stouffer et al., 1949; Nguyen et al., 2016c), modern effect-size-based approaches (Haynes et al., 2017) and rank aggregation-based approaches (Pihur et al., 2009). However, these approaches may not be suitable for revealing the mechanism of action for a given disease since they do not account for the heterogeneity that is present across multiple data types (mRNA, miRNA, DNA methylation, etc.). The approaches in the second category combine sample-matched studies from multiple data types and provide biomarkers that can capture data heterogeneity present across the *omic* layers. Integrating such information from multiple data types is essential for obtaining a comprehensive overview of the given biological system and thought to provide better prognostic markers (Berger et al., 2013; Kristensen et al., 2014; Nguyen et al., 2016b). For instance, it has been shown that integrating miRNA and mRNA expression profiles results in greater statistical power and better understanding of the underlying disease phenomena, both in the context of biomarker discovery (Volinia and Croce, 2013; Wotschofsky et al., 2016) and pathway analysis (Calura et al., 2014; Vlachos et al., 2015; Alaimo et al., 2016; Diaz et al., 2016). More recently, it has been demonstrated that the integration of long non-coding RNA (lncRNA) and mRNA plays an important role in revealing pathogenetic mechanisms of a given condition (Lin et al., 2014; Liu et al., 2018). However, these approaches require the same group of individuals to be present for each of the experiments coming from different *omic* layers. Thus, they fail to utilize the information from dozens of independent studies containing thousands of samples for a given disease that is currently available in public repositories such as Gene Expression Omnibus (GEO) (Barrett et al., 2005), TCGA [http://cancergenome.nih.gov] or ArrayExpress (Rustici et al., 2013).

DNA methylation has been recognized to play a crucial role in cancer progression (Esteller, 2008; Parrella, 2010). An increasing number of computational approaches have been published in recent years for the identification of methylation-based biomarkers (Gevaert et al., 2015; Hao et al., 2017; Hong et al., 2017; Shafi et al., 2018). However, to the best of our knowledge, none of the current approaches is able to identify network-based gene signatures considering the data heterogeneity among the independent DNA methylation and gene expression studies. The approach presented in this manuscript bridges this gap.

Here we propose a multi-cohort and multi-omics meta-analysis framework that is able to integrate unmatched mRNA and DNA methylation data obtained from many different independent studies, and subsequently identify network-based signatures that can capture putative mechanisms of a given disease. We apply our proposed framework on nine independent datasets related to glioblastoma (GBM) containing a total of 622 samples and eight independent studies related to low-grade glioma (LGG) containing a total of 1,787 samples. The identified network-based signatures are validated based on their ability to predict the patients' clinical outcome for 1,269 samples from four completely independent validation datasets. This is done by clustering the patients included in the validation datasets using perturbation clustering (Nguyen et al., 2017b), which identifies the correct number of clusters present in the data and groups the patients accordingly. The signatures extracted from the proposed framework are then compared with 10 other previously published gene signature panels related to GBM and LGG. For both diseases, the network-based signatures identified by our proposed framework are able to separate patients associated with poor survival from other individuals with significant Cox *p*-values and outperform the other compared signatures. This suggests that the proposed framework is able to provide better prognostic biomarkers compared to the existing ones.

2. MATERIALS AND METHODS

The goal of the proposed framework is to identify reliable network-based gene signatures by integrating independent

experiments obtained from multiple data types. The framework takes three types of inputs: (i) mRNA datasets, (ii) DNA methylation datasets, and (iii) known gene interaction networks. The mRNA and DNA methylation datasets can be completely independent, which means that they can be obtained from different experiments performed in different laboratories and can include samples from different cohorts of patients. The gene interaction network is a graph in which the nodes represent genes and the edges represent interactions between them. This information can be obtained from any resources that describe the known gene-gene interactions such as KEGG, Reactome, STRING, or HPRD.

Each mRNA or methylation dataset is represented by a matrix in which the rows represent the measured genes and the columns represent the samples included in the given study. The value in each cell reflects the measured expression or methylation level of a gene for a particular sample. Each dataset includes samples from two given phenotypes such as disease vs. healthy, treated vs. non-treated, disease subtype A vs. disease subtype B, etc.

The overall workflow of the proposed framework is divided into four main modules (**Figure 1**). The first two modules, described in section 2.1, account for the variability across the individual datasets coming from the same data type, while the third and fourth modules, described in section 2.2, account for the variability across the data types (mRNA and methylation) and integrate network information into the framework in order to identify impacted subnetworks. Briefly, the first module takes the given list of mRNA datasets as input and performs a meta-analysis to identify the genes that are differentially expressed across the given phenotypes. Due to the heterogeneity present in the individual mRNA datasets, the identified list of genes might be significantly impacted by a single study, and hence might not represent the true list of genes impacted for the given condition. Therefore, a *leave-one-out* (Friedman et al., 2001) meta-analysis is carried out to make the list of genes more reliable. The second module takes the given list of methylation datasets as input and utilizes the same meta-analysis pipeline to identify the genes that are differentially methylated across the given phenotypes. The third module combines the results obtained from the first two modules and identifies the genes that are driven by their methylation profiles. This module essentially integrates information obtained from two *omic* layers (transcriptomic and epigenomic) and takes into account the heterogeneity that may arise across these layers. Finally, the fourth module incorporates the known interactions among the genes and identifies the subnetworks that are affected by the methylation-driven genes.

2.1. Multi-Cohort Meta-Analysis

This section describes the first and second modules of the framework (**Figures 1A,B**). The meta-analysis pipeline proposed here utilizes both classical *p*-value-based and modern effect-size-based meta-analysis to calculate gene level statistics. The backbone of this algorithm is an extended version of the meta-analysis framework proposed in one of our previously published works (Nguyen et al., 2016a). The overall pipeline consists of three steps: (i) obtaining *p*-values from classical hypothesis testing, (ii) obtaining effect sizes and their *p*-values and (iii) combining the two types of *p*-values to calculate the final gene level statistics. The first two steps are independent of each other and can be performed concurrently.

At first, two-tailed *p*-values are calculated for all genes across all studies by performing a classical hypothesis testing. A moderated *t*-test provided by limma (Smyth, 2005) is utilized for this purpose. This can also be replaced with other classical tests such as two sample *t*-test, paired *t*-test, etc.

If the input matrix contains discrete values (e.g., data obtained from RNA-seq experiment or bisulfite sequencing experiment), regression-based approaches such as Poisson, quasi-Poisson or negative binomial regression models should be used instead (Robinson et al., 2010; Anders et al., 2012; Klein and Hebestreit, 2015; Shafi et al., 2018). The two-tailed *p*-values are then converted to one-tailed (left- and right-tailed) *p*-values. Gene level *p*-values generated by the individual studies are then combined by using *addCLT* (Nguyen et al., 2017a), an *additive approach* (Edgington, 1972) based on the *Central Limit Theorem* (Kallenberg, 2002) that is robust against outliers. For each gene, this *p*-value represents the chance of observing its combined differential expression (or methylation) just by chance.

To estimate the effect size, we first calculate the standardized mean difference (SMD) of each gene across all studies. Considering SMD instead of the raw mean difference is crucial since the expression (or methylation) levels within each study might be scaled differently. In this work, we use Hedge's *g* (Hedges and Olkin, 2014) as the SMD to measure expression (or methylation) changes between the two given phenotypes. Central tendencies for the effect sizes are calculated using the random-effect model and the REstricted Maximum Likelihood (REML) algorithm (Viechtbauer, 2010). Next, we calculate the z-scores and left- and right-tailed *p*-values of the z-scores to estimate the probability of observing such effect sizes just by chance. This overall estimated effect size represents the expression (or methylation) change of a gene under the effect of the given condition.

In the third step, we combine the two types of evidence (one obtained from classical hypothesis testing, another from estimating the effect sizes) using a conservative maxP (Wilkinson, 1951) method. We are using this conservative statistic because we want a significant *p*-value only if the gene is significant based on both classical *p*-value-based and the more modern effect-size-based meta-analysis. The *p*-values are corrected for multiple comparisons using an FDR approach. Finally, a predefined threshold is used to select the genes that are differentially expressed or methylated.

2.2. Multi-Omics Data Integration

This section describes the third and fourth modules of the framework. The inputs of the third module (**Figure 1C**) are two lists of genes obtained from the meta-analysis step described in section 2.1 above. The first list includes the differentially expressed genes (DEGs), while the second one includes the differentially methylated genes (DMGs) across the given phenotypes. From these two lists of genes, we first select the genes that are present in both lists, i.e., the genes that are both differentially expressed and methylated. Next, we filter them by

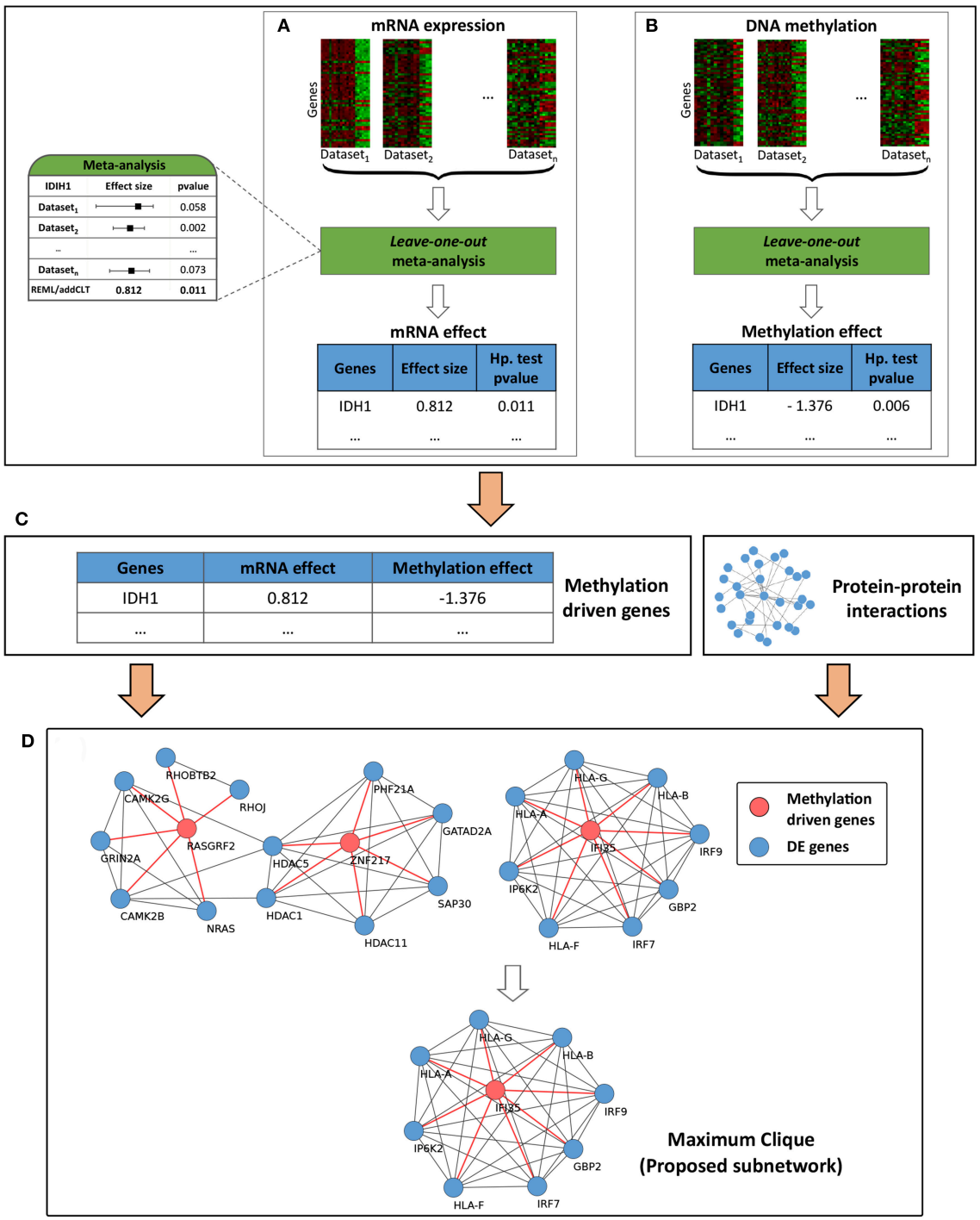


FIGURE 1 | Overall workflow of the proposed framework. **Module (A)** takes multiple independent mRNA datasets and performs a leave-one-out meta-analysis to identify reliable differentially expressed genes (DEGs). Similarly, **module (B)** takes multiple independent DNA methylation datasets and identifies differentially methylated genes (DMGs). DEGs and DMGs are then systematically integrated in **module (C)** to identify methylation-driven genes (MDGs). Finally in **module (D)**, the MDGs are used as inputs in a network propagation algorithm to identify the proposed subnetworks.

selecting the genes for which the mRNA and methylation changes occurred in opposite directions. This is motivated by the fact that methylation correlates negatively with gene expression (Shafi et al., 2018). In other words, when a CpG site is methylated in the promoter regions, it typically represses the transcriptional activity of that region by restricting the binding of specific transcription factors (TFs). Alternatively, when a CpG site is unmethylated in the promoter regions, it allows for the binding

of those TFs (Jones, 2012). Finally, we identify the methylation-driven genes (MDGs) by filtering the genes that have unsigned effect sizes lower than a given threshold. This is an optional step of the framework. The default threshold is set to zero (no filtering).

Identified MDGs can be thought of as individual gene markers that can distinguish the phenotypes of a given disease, based on both individual mRNA and methylation data. However, to better understand the underlying disease mechanisms, and to better predict patient prognosis, it is important to incorporate known information about the interactions between the genes (Mitra et al., 2013).

The fourth module of the framework (**Figure 1D**) uses the identified MDGs, DEGs and the given network information to identify the subnetworks that are perturbed by the signals propagated through the edges of the MDGs. For each MDG, we create its own DE neighborhood by selecting the DEGs that are directly connected with it. All identified subnetworks are then merged together into a larger network. This concept of network propagation has been used by several research groups for active subnetwork identification using transcriptomic data (Komurov et al., 2012; Ansari et al., 2017) and mutational hotspot identification in human cancers (Ciriello et al., 2012). Finally, within this larger network, we select the genes that are part of the largest *cliques* as our proposed signature. This idea is driven by the fact that cliques are fully-connected subnetworks in which all nodes are connected in a pairwise fashion; and therefore, genes that are part of a clique are more likely to be functionally related (Pradhan et al., 2012).

2.3. Perturbation Clustering

In order to evaluate the prognostic value of the proposed signature, we use the genes present in the signature to identify disease subtypes from the independent patient cohort. For clustering, we use PINS (Nguyen et al., 2017b; Nguyen H. et al., 2018) to perform perturbation clustering that was developed in our research lab for tumor subtyping. PINS can automatically determine the number of clusters and then identify subtypes that are the most stable against noise and data perturbation. PINS is developed based on the observation that small changes in any kind of quantitative assay will be inherently present between individuals, even in a truly homogeneous population in the absence of any molecular subtypes. Therefore, well-defined subtypes of a disease have to be stable with respect to small changes in the measured values. In order to identify robust subtypes, PINS repeatedly perturbs the data by adding Gaussian noise and then clusters the patients. PINS yields subtypes and patient patterns that are least affected by data perturbation. More details of the algorithm can be found in Nguyen et al. (2017b).

Here, the input of the subtyping algorithm is a matrix in which the rows represent the patients and the columns represent the signature genes identified by our framework. Different gene signatures yield different matrices (same set of patients/rows but different sets of genes/columns). We expect that a better signature will provide better subtyping, i.e., subtypes with more significant survival differences. The number of clusters (k) is automatically determined by PINS. We simply used the default settings of the PINS R package (Nguyen H. et al., 2018).

3. RESULTS

We demonstrate the performance of the proposed framework by constructing network-based signatures for two diseases: glioblastoma multiforme (GBM) and low-grade glioma (LGG). In the GBM study, we included only the stage IV glioma tumors, whereas in the LGG study we included stage II and III glioma tumors. This is consistent with others such as TCGA (Cancer Genome Atlas Research Network et al., 2015), Noushmehr et al. (2010) and Garkavtsev et al. (2004), who also considered stage II and III glioma tumors as LGG. All staging is based on the World Health Organization (WHO) standard. All discovery datasets used in this manuscript were obtained from GEO (Barrett et al., 2005). Dataset summaries and preprocessing techniques are described in the **Supplementary Materials**. We downloaded the protein-protein interaction (PPI) networks from the STRING database version 10.5 to obtain information about the gene interactions. STRING provides a confidence score (ranging from 0 to 1,000) for each interaction in the network. Here we used a score of 900 to select the high confidence interactions, resulting in a network of 9,941 genes and 227,186 interactions (top 4.9% interactions).

One of the most widely accepted techniques to evaluate the prognostic performance of a gene signature is to test its ability to predict patients' survival in independent datasets (Chang et al., 2005; Shedden et al., 2008; Szász et al., 2016). In order to achieve this goal, we used PINS (described in section 2.3) on independent gene expression validation datasets obtained from three different sources: (i) TCGA, (ii) GEO, and (iii) CGGA (Yan et al., 2012; Sun et al., 2014). None of these datasets have been used in the original training datasets. PINS can automatically determine the number of clusters (denoted by k). We use only the list of genes present in the proposed subnetwork as features, instead of all genes present in the datasets. Survival analysis is performed using Kaplan–Meier survival analysis (Kaplan and Meier, 1958) and their statistical significance is assessed using a Cox regression model (Cox, 1972).

3.1. Glioblastoma (GBM) Study

We first identify 2,183 DEGs by performing *leave-one-out* meta-analysis (section 2.1) on four mRNA datasets (GSE7696, GSE4290, GSE90598, and GSE22866). Similarly, we analyze five methylation datasets (GSE60274, GSE22867, GSE50923, GSE79122, and GSE36278) and identify 1,205 DMGs. These nine discovery datasets include a total of 622 samples: 533 samples from GBM patients and 89 from healthy (non-tumor) individuals. Descriptions of these datasets are provided in **Table S1**. We use a stringent threshold of 0.1% for both differential expression and methylation.

Next, we identify the list of methylation-driven genes (MDGs) based on the three following criteria: (i) genes present in the list of DEGs with absolute mRNA effect sizes $>$ 1, (ii) genes present in the list of DMGs with absolute methylation effect sizes $>$ 1, and (iii) genes that have opposite mRNA

and methylation effect sizes (i.e., genes with positive mRNA effect sizes need to have negative methylation effect sizes, while genes with negative mRNA effect sizes need to have positive methylation effect sizes). The identified list contains 45 MDGs. Each of these identified MDGs are then used as seeds in the network propagation step to build neighbor networks of DEGs (section 2.2). These subnetworks are then merged together to form a larger network, containing a total of 214 candidate genes. Finally, within the larger network, the largest cliques contain 46 genes which constitute the proposed network-based signature for this disease (**Figure 2**).

We demonstrate the utility of the proposed signature on two independent gene expression datasets; one, downloaded from the TCGA GBM cancer site (The Cancer Genome Atlas Research Network, 2013), contains gene expression profiles of 525 individual patients, and the other one, GSE4412 (Freije et al., 2004), was downloaded from GEO and contains gene expression profiles of 59 individual patients. For both datasets, our proposed signature combined with PINS is able to identify two groups of patients with significantly different survival rates using the Cox regression model. The Cox *p*-value for TCGA datasets is 7.38E-04, whereas the Cox *p*-value for GSE4412 is 9.70E-03.

We compare our signature with the following 7 previously published GBM gene signature panels: 9 methylation-based gene signature proposed by Shukla et al. (2013), 13 methylation-based gene signature proposed by Etcheverry et al. (2010), 14 prognostic gene signature proposed by Arimappamagan et al. (2013), 35 methylation based gene signature proposed by Smith et al. (2014), 35 prognostic gene signature proposed by Fatai and Gamieldien (2018), 36 methylation-based gene signature proposed by Chiang et al. (2014) and 48 gene signature proposed by Crisman et al. (2016).

The comparison based on the prognostic performances of these gene signature panels is shown in **Table 1**. Related survival curves are shown in **Figure 3**. PINS identifies the optimal number of clusters based on the given input, which is denoted by *k* in the table. The cells highlighted in yellow represent the Cox *p*-values that are significant (< 0.01). The cells highlighted in green show the best signature (i.e., lowest Cox *p*-value) for each dataset. These results show that in both datasets, the proposed signature achieves the best results. Furthermore, in the GSE4412 dataset, only the proposed signature is able to achieve a significant Cox *p*-value.

3.2. Low-Grade Glioma (LGG) Study

Similar to the previous study, here we perform *leave-one-out* meta-analysis on five mRNA datasets (GSE16011_cohort1, GSE16011_cohort2, GSE4290, GSE68848, and GSE4271) and three DNA methylation datasets (GSE90496, GSE109379, and GSE53227), and identify 1,564 DEGs and 2,721 DMGs

TABLE 1 | Prognostic performance of different gene signature panels related to GBM.

		TCGA (525 patients)		GSE4412 (59 patients)	
Gene signatures	**Number of genes**	***k***	**Cox *p*-value**	***k***	**Cox *p*-value**
Proposed signature	46	2	7.38E-04	2	9.70E-03
Shukla et al.	9	5	3.76E-03	5	1.12E-02
Etcheverry et al.	13	5	3.42E-03	3	7.50E-01
Arimappamagan et al.	14	2	3.14E-03	5	4.67E-01
Smith et al.	35	3	9.26E-03	3	6.07E-01
Fatai et al.	35	3	1.01E-01	3	3.93E-01
Chiang et al.	36	4	8.88E-01	4	9.98E-02
Crisman et al.	48	5	3.61E-02	5	4.17E-01

Clustering is performed by using PINS. The number of clusters identified by the algorithm is denoted by k. The cells highlighted in yellow represent the Cox p-values that are significant (<0.01). The cells highlighted in green represent the best signature (i.e., lowest Cox p-value) for each dataset. These results indicate that the proposed signature is able to achieve the lowest Cox p-values on both independent datasets.

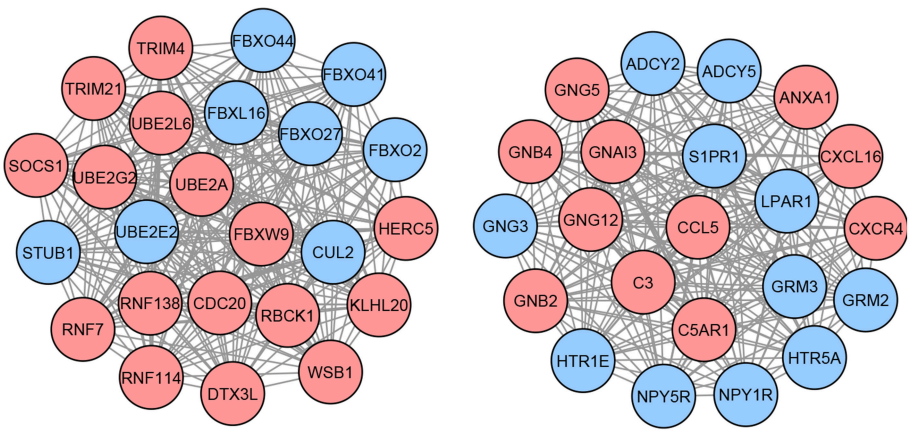


FIGURE 2 | Proposed network-based signature for GBM, containing a total of 46 genes organized in two different cliques. Each node in this graph represents a gene, while each edge describes the interaction between a gene pair. The interactions are retrieved from the STRING database. The colors of the nodes represent the effect sizes obtained from the meta-analysis step described in **Figure 1A**: red represents genes with a positive effect size while blue represents genes with a negative effect size.

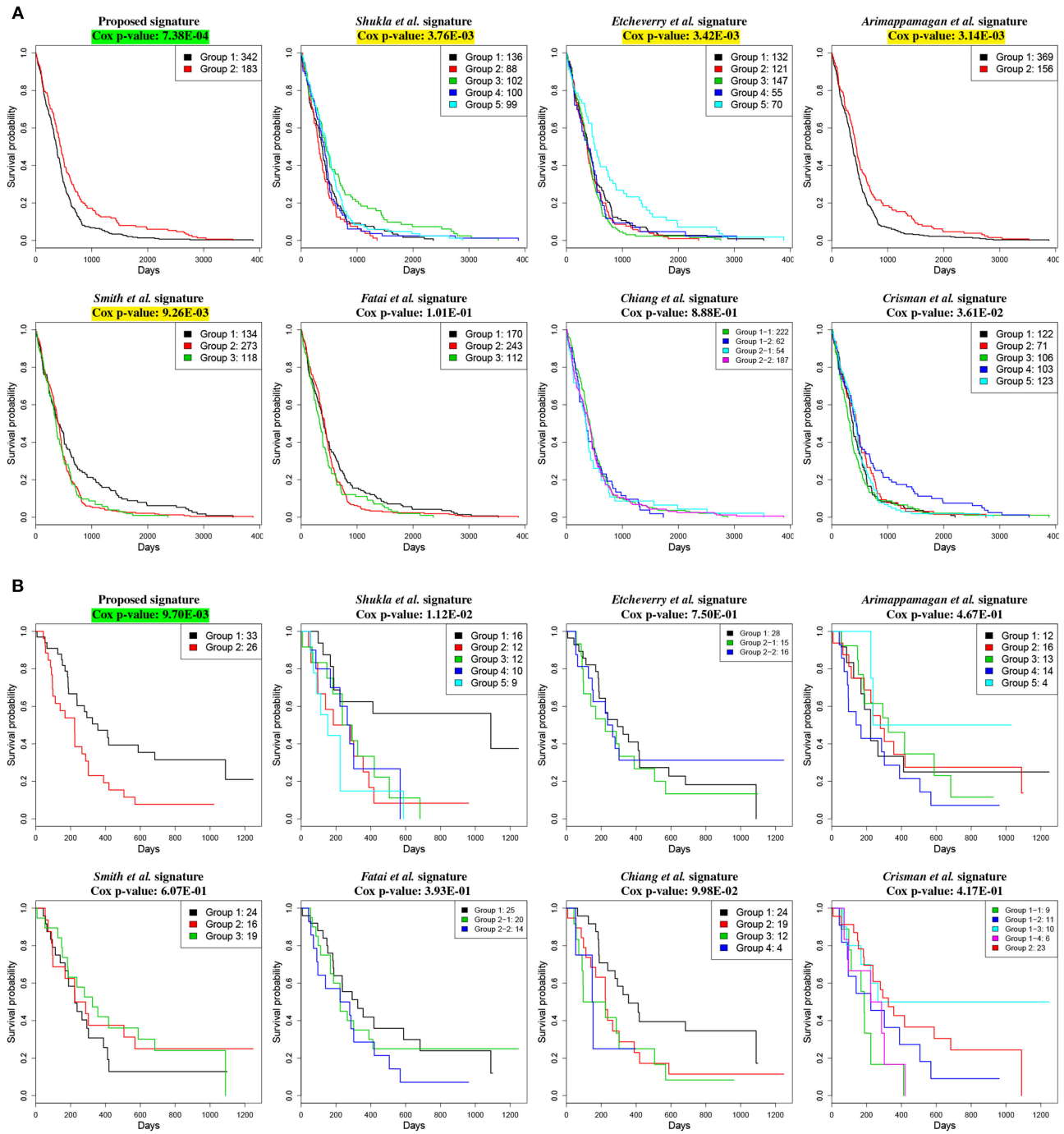


FIGURE 3 | Kaplan–Meier survival analysis on GBM studies, using different gene signature panels. **(A)** TCGA dataset which contains gene expression profiles from 525 individual patients. **(B)** GSE4211 dataset which contains gene expression profiles from 59 individual patients. The horizontal axes represent the time (in days) from the start of the study, whereas the vertical axes represent estimated survival percentage. Yellow colors represent the Cox *p*-values that are significant (<0.01). The green color indicates the best signature (i.e., lowest Cox *p*-value) for the given dataset. These results show that the proposed signature yields the best separation between aggressive and less aggressive disease on both datasets.

respectively. These eight datasets contain a total of 1,787 samples. Among them, 1,026 samples are from LGG patients while 761 from either GBM patients or healthy (non-tumor) individuals. Descriptions of these datasets are provided in **Table S2**. In this study, we use a threshold of 5% for differential expression and methylation.

After integrating DEGs and DMGs in the third module, we find 52 methylation-driven genes (MDGs). Next, we perform network propagation to construct the subnetworks that contain the DEGs directly connecting to MDGs. After merging these subnetworks, we obtain a list of 110 candidate genes. Finally, 20 genes are selected based on the maximum clique present in the network which is the proposed signature for this study. The identified network-based signature is shown in **Figure 4**.

To demonstrate the utility of the proposed signature, we use two independent gene expression datasets; one from TCGA LGG cancer site (Cancer Genome Atlas Research Network et al., 2015) that contains a total of 515 patients, and the other one from CGGA that contains a total of 170 patients. We use PINS to perform a perturbation clustering using the genes present in the proposed network as features. Similar to the GBM study, for both datasets, the groups of patients identified based on the given signature have significantly different survival profiles. For the TCGA dataset, the Cox *p*-value is 5.48E-09 with 4 clusters whereas for the CGGA dataset, the Cox *p*-value is 1.82E-04 with 5 clusters.

We compare our proposed signature with the following 3 published LGG gene signature panels: a set of 6 genes identified by Olar and Sulman (2015), a meta-signature of 20 genes proposed by Wang et al. (2017) and a panel of 24 genes proposed by Liu et al. (2011). The comparison between the results obtained with these signatures is shown in **Table 2**. The related survival curves are shown in **Figure 5**. In the TCGA dataset, the proposed signature and the signature proposed by Liu et al. achieve significant Cox *p*-values. In CGGA dataset, significant Cox *p*-values are achieved by the proposed signature and the signature proposed by Olar et al. These results show that in both datasets, the proposed signature achieves the best results.

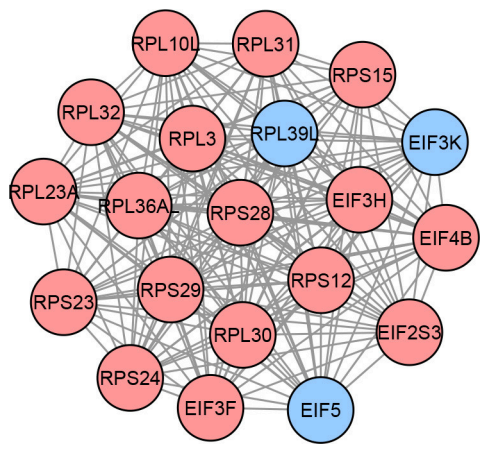


FIGURE 4 | Proposed network-based signature for LGG, containing a total of 20 genes organized in a clique. Each node in this graph represents a gene, while each edge describes the interaction between a gene pair. The interactions are retrieved from the STRING database. The colors of the nodes represent the effect sizes obtained from the meta-analysis step described in **Figure 1A**: red represents genes with a positive effect size while blue represents genes with a negative effect size.

3.3. Network-Based Signature vs. Methylation-Driven Genes (MDGs)

To demonstrate the contribution of the network information in our framework, we compare the prognostic performance of the proposed network-based signature with the performance of a signature derived from methylation-driven genes (MDGs) alone. **Table 3** shows the Cox *p*-values obtained by using these two types of signatures on the four independent datasets used in the above two studies. PINS was used to group the samples. For GBM, the MDGs and the proposed signature contain 45 and 46 genes respectively, while for LGG, the MDGs and the proposed signature contain 27 and 20 genes, respectively. Results indicate that, for both diseases (each disease contains two independent datasets), network-based signatures outperform the individual markers (i.e., MDGs) based on their ability to predict the patients' clinical outcome.

4. DISCUSSION

One widely used technique to combine multiple independent studies is to perform a *horizontal* meta-analysis (i.e., combining sample-unmatched studies of the same data type). This approach is unable to combine studies coming from multiple data types. Hence, it is not suitable for the identification of the *mechanism of action* of a given disease. Another technique is to perform a *vertical* meta-analysis (i.e., combining sample-matched studies from multiple data type) which accounts for the heterogeneity that may arise across different *omic* layers. However, the latter technique requires each data type to be available for each individual patient, which is expensive and impractical for the studies with large sample sizes. To overcome these challenges, in this manuscript, we propose a multi-cohort and multi-omics meta-analysis framework that identifies network-based signatures using independent mRNA and DNA methylation studies available in the public repositories. The identified signatures are evaluated based on their ability to distinguish

TABLE 2 | Prognostic performance of different gene signature panels related to LGG.

		TCGA (515 patients)		CGGA (170 patients)	
Gene signatures	**Numer of genes**	***k***	**Cox *p*-value**	***k***	**Cox *p*-value**
Proposed signature	20	4	5.48E-09	5	1.82E-04
Olar et al.	6	5	6.97E-02	5	5.43E-03
Wang et al.	20	2	1.42E-01	4	8.07E-01
Liu et al.	18	5	3.21E-06	2	1.12E-02

Clustering is performed by using PINS. The number of clusters identified by the algorithm is denoted by k. The cells highlighted in yellow represent the Cox p-values that are significant (<0.01). The cells highlighted in green represent the best signature (i.e., lowest Cox p-value) for each dataset. These results indicate that the proposed signature is able to achieve the lowest Cox p-values on both independent datasets.

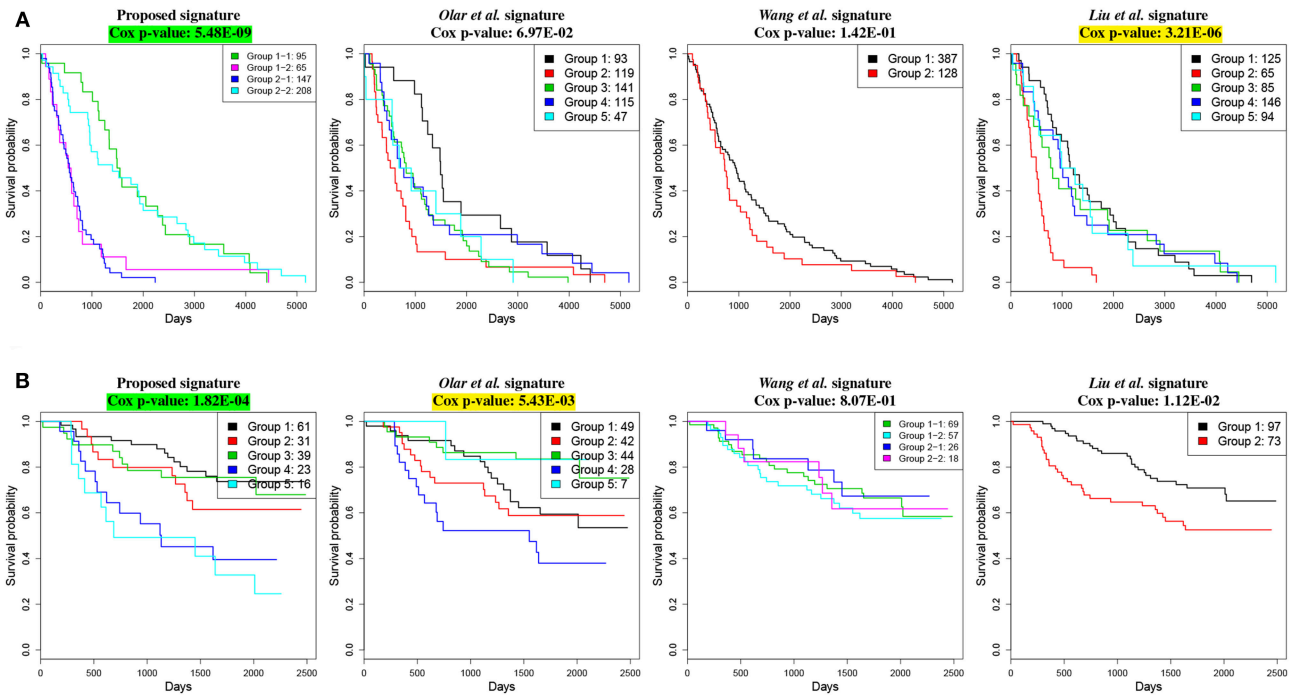


FIGURE 5 | Kaplan–Meier survival analysis on LGG studies, using different gene signature panels. **(A)** TCGA dataset which contains gene expression profiles from 515 individual patients. **(B)** CGGA dataset which contains gene expression profiles from 170 individual patients. The horizontal axes represent the time (in days) from the start of the study, whereas the vertical axes represent estimated survival percentage. Yellow colors represent the Cox *p*-values that are significant (<0.01). The green color indicates the best signature (i.e., lowest Cox *p*-value) for the given dataset. These results show that the proposed signature yields the best separation between aggressive and less aggressive disease on both datasets.

TABLE 3 | Prognostic performance of network-based signatures vs. individual markers.

	GBM study					LGG study				
		TCGA GBM		GSE4412			TCGA LGG		CGGA	
Gene signatures	***m***	***k***	**Cox P**	***k***	**Cox P**	***m***	***k***	**Cox P**	***k***	**Cox P**
Meth. driven genes (MDGs)	45	4	9.36E-03	3	1.18E-01	27	3	3.22E-06	2	1.43E-03
Network-based signature	46	2	7.38E-04	2	9.70E-03	20	4	5.48E-09	5	1.82E-04

Clustering is performed by using PINS. Number of clusters identified for a given dataset is denoted by k, while the number of genes for a given study is denoted by m. Cells highlighted in green represent the best signature (i.e., lowest Cox p*-value) for each dataset. Results indicate that incorporating network information leads to better prognostic gene markers.*

patients with different survival profiles on independent validation datasets.

One of the inputs required for the proposed framework is the known interactions between the genes. This information can come from any protein-protein interaction database for the given organism and is independent of the specific experiment or condition. In our case, this type of data came from the STRING database, which would be suitable for any experiment involving more than 2,000 organisms. The discovery datasets used in this manuscript are downloaded from GEO. We have included all gene expression and methylation studies related to GBM and LGG that have a total number of samples measuring 20 or more after data preprocessing. Datasets from any other resources such as TCGA, ArrayExpress (Rustici et al., 2013), etc., can also be used as long as they contain samples from two phenotypes (disease vs. normal, treated vs. non-treated, etc.). The framework is appropriate for the disease conditions whose mechanisms of actions are known to be triggered by the change in DNA methylation. Due to the important role of DNA methylation in glioma (Heyn and Esteller, 2012; Turcan et al., 2012), we demonstrate our proposed framework on two subtypes of glioma; the most aggressive one, GBM, and the comparatively less aggressive LGG. However, this framework can be used to identify network-based markers for other disease conditions as well.

We leverage the concept of the network propagation algorithms mentioned in Mitra et al. (2013) to identify candidate subnetworks from the methylation-driven genes.

The final network-based markers are selected based on the maximum clique. Cliques are complete graphs in which all nodes are connected in a pairwise fashion, and therefore, genes that are part of a clique are likely to be functionally related. In previous years, the utility of using cliques has been demonstrated in multiple disease conditions such as breast cancer (Shi et al., 2010), colorectal cancer (Pradhan et al., 2012), etc. Other subnetwork identification techniques, such as greedy algorithms (e.g., PinnacleZ, Chuang et al., 2007), clustering-based methods (e.g., SAMBA, Tanay et al., 2004), scoring based on centrality measurements (e.g., Wang et al., 2011), etc., can be utilized as well. A comprehensive review of the currently available tools for subnetwork identification can be found in Nguyen et al. (2019).

We investigate how the groups of patients identified in the TCGA GBM dataset, using our proposed signature (**Figure 3A**), relate with the available histopathological variables or treatments. **Table S3** shows the confusion matrix of the two groups of patients associated with the proposed GBM signature and the five GBM subtypes recognized by the original authors (The Cancer Genome Atlas Research Network, 2013). Enrichment analysis using Fisher's Exact Test (FET) indicates that the group of patients with lower survival rate is enriched with Mesenchymal subtype (p = 1.04E-19), whereas the group of patients with higher survival rate is associated with Proneural (p = 1.98E-14) subtype and G-CIMP tumors (p = 4.27E-10). This confirms the fact that G-CIMP tumors belong to the Proneural subtype (Noushmehr et al., 2010; Verhaak et al., 2010). In addition, the better survival group is enriched with IDH1 mutation (p = 1.80E-06) and relatively younger patients (Wilcoxon rank sum (WRS) test p = 0.01), which is also acknowledged by others (Noushmehr et al., 2010; The Cancer Genome Atlas Research Network, 2013). Furthermore, we investigate patients' responses to Temozolomide (TMZ), a drug which is FDA approved for the treatment of GBM. We do this by calculating the survival Cox p-value for each group (the better survival group and the lower survival group) based on the patients treated with and without TMZ (treated with other drugs or untreated). The results indicate that only one group of patients (not both) is associated with favorable TMZ drug response, which is reflected by significantly different survival rates of the drug-responders and the drug-resistants (Cox p-value = 7.34E-06). Our finding explains why it has previously been noted that there is a group of patients who do not respond well to TMZ (Kitange et al., 2009; Lee, 2016).

Similarly, to investigate the groups of patients identified on TCGA LGG, we obtained clinical information from TCGA that includes three subtypes of glioma: IDH wild-type, IDH mutant-codel, and IDH-mutant-non-codel (Ceccarelli et al., 2016). Enrichment analysis using FET reveals that the groups of patients with lower survival rates (cluster "1-2" and "2-1" in **Figure 5A**) are enriched with wild-type IDH (p = 2.30E-16 and 1.94E-06) and MGMT promoter unmethylation (p = 4.99E-06 and 0.001). These results confirm the findings previously reported by TCGA and others (Hegi et al., 2005). In addition, we found that the lower survival rates are associated with a higher tumor purity score (WRS p-value = 0.007). Previously, it has been shown by others that a higher tumor purity score is associated with tumor growth, disease progression and drug resistance (Yoshihara et al., 2013).

We also investigate the novelty of our identified signatures by checking their overlap with other published signature genes (**Figure 6**). For GBM, none of the genes proposed in this manuscript are present in the other three top (based on the Cox p-value on TCGA dataset) gene signature panels (i.e., panels of gene signatures proposed by Shukla et al., Etcheverry et al., and Arimappamagan et al.). Similarly for LGG, none of the genes proposed in this manuscript are present in the panels of gene signatures proposed by Olar et al., Wang et al., and Liu et al.

One of the main reasons for this is that the types of evidence used by our proposed framework are different from other relevant studies. Our proposed framework identifies gene signatures using evidence from three different sources: (i) mRNA expression, (ii) DNA methylation, and (iii) protein-protein

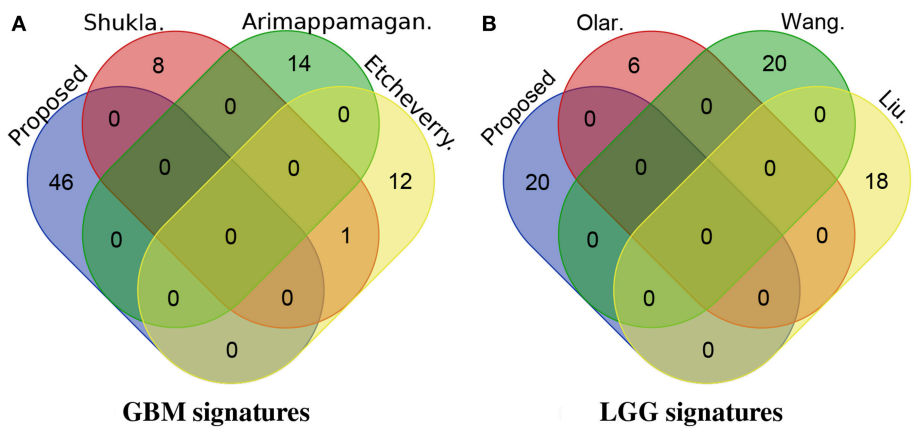
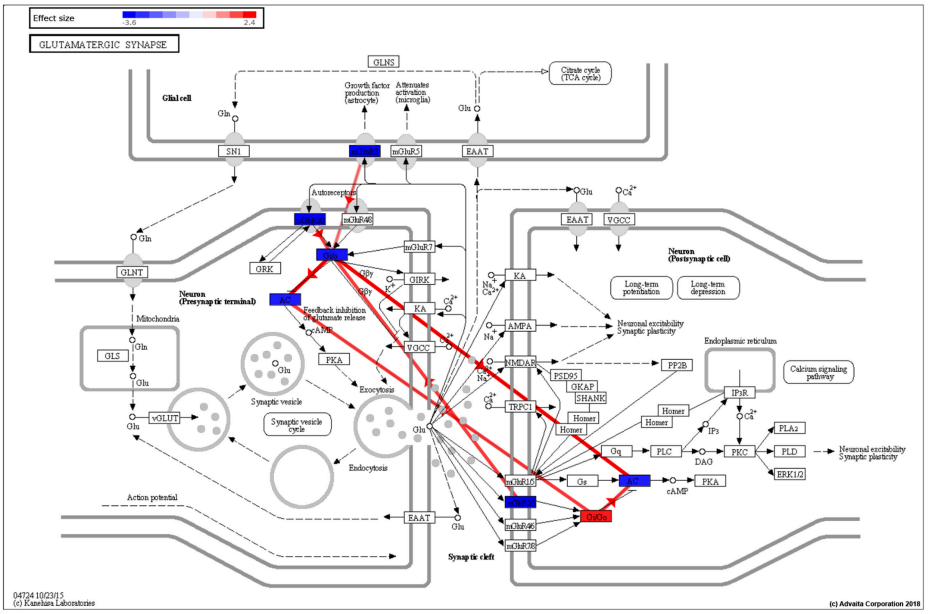


FIGURE 6 | The overlap between the proposed signatures and other previously published signatures. For GBM, none of the genes proposed in this manuscript are present in the signatures proposed by Shukla et al., Etcheverry et al., and Arimappamagan et al. Similarly for LGG, none of the genes proposed in this manuscript are present in the signatures proposed by Olar et al., Wang et al., and Liu et al. **(A)** GBM signatures. **(B)** LGG signatures.

interactions (PPI). In addition, it combines heterogeneous independent studies within each data type (mRNA and DNA methylation) using an effect-size-based meta-analysis approach. In contrast, none of the relevant studies identify their gene signatures considering all three types of evidence that we used. They are based on frameworks that either do not integrate information from multiple data levels or do not combine multiple studies within one data level, or both. Therefore, a very small or no overlap between the signatures proposed by our framework and the signatures proposed by other relevant studies is to be expected. Furthermore, the existing signatures have little or no overlap among themselves, even though many of them are based on the same type of evidence. In spite of the fact that our proposed genes have not been previously

Glutamatergic synapse pathway



Chemokine signaling pathway

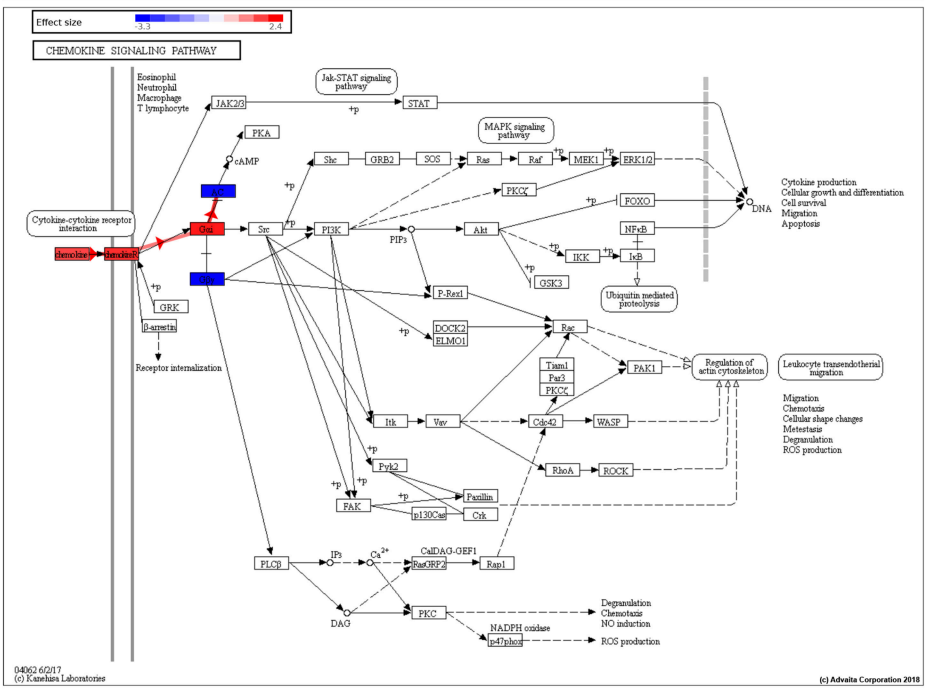


FIGURE 7 | Interesting putative mechanisms are identified by iPathwayGuide (www.advaitabio.com) on the **Glutamatergic synapse** and the **Chemokine signaling** pathways. The colors of the nodes represent the effect sizes obtained from the meta-analysis step described in **Figure 1A** of the manuscript: red represents genes with a positive effect size while blue represents genes with a negative effect size. The edges highlighted in red represent the *coherent edges* between the genes, which indicate the edges for which the measured effect changes are consistent with the phenomena described by the pathway.

reported, they provide the best ability to distinguish between aggressive and less aggressive disease in all independent datasets that we used.

Importantly, our proposed GBM signature contains several genes that play crucial roles in the underlying mechanisms of GBM. For instance, according to Deng et al. (2016), ADCY2 is known to be involved in the progression of diffuse intrinsic pontine glioma; ANXA1 has shown to be involved in GBM apoptosis by Festa et al. (2013); Pan et al. (2017) demonstrated that CCL5 is responsible for creating an autocrine circuit for Mesenchymal GBM growth; Xie et al. (2015) investigated the role of CSC20 and found its crucial role in tumor-initiating cell (TIC) proliferation in GBM; CXCR4, LPAR1 and TRIM21 play important roles GBM cell proliferation as demonstrated by Ehtesham et al. (2009), Loskutov et al. (2018), and Lee et al. (2017), respectively; Kim et al. (2018) demonstrated the therapeutic role of RNF138 in GBM; Mahajan-Thakur et al. (2017) reviewed the role of S1PR1 in GBM and found that its over-expression is associated with improved GBM prognosis; SOCS1 plays a vital role as a tumor suppressor in GBM, as investigated by Baker et al. (2009); STUB1 has shown to be involved in glioma cell proliferation by Syed et al. (2015); etc. Similarly, our proposed LGG signature contains genes that are known to be related to glioma. For instance, according to Shi et al. (2006), EIF3F is downregulated in most human tumors including glioma; EIF5 and RPS12 are known to be involved in brain metastasis in primary breast tumors (Sanz-Pamplona et al., 2011); Shahbazian et al. (2010) has shown that EIF4B is a potential target for anti-cancer therapies; etc.

Furthermore, we use iPathwayGuide (Advaita Corporation, 2019) to perform an extensive pathway analysis to identify the mechanisms captured by the proposed signatures. iPathwayGuide uses an impact analysis that calculates the true impact of a pathway by combining two types of evidence. The first type of evidence is the classical over-representation of DE genes in each pathway. The second type of evidence captures several other important biological factors such as the position of all the genes on each pathway, the magnitude of their expression change, the direction and type of the signals transmitted between genes as described by the pathway, etc. The impact analysis has been shown to be able to identify the significantly impacted pathways much better than classical over-representation alone (Drăghici et al., 2007; Tarca et al., 2009).

Among the pathways reported as significant, interesting putative mechanisms are identified by the impact analysis on the Glutamatergic synapse pathway and the Chemokine signaling pathway. These are shown in **Figure 7**. The colors of the nodes represent the effect sizes obtained from the meta-analysis step described in **Figure 1A**: red represents genes with a positive effect size while blue represents genes with a negative effect size. The edges highlighted in red represent *coherent edges*. A coherent edge is an edge for which the measured effect changes are consistent with the phenomena described by the pathway. For example, if gene A inhibits gene B, and if gene A is upregulated, gene B is expected to be downregulated. If the measured changes are consistent with this inhibition, the edge corresponding to this interaction is referred to as being coherent. Several such coherent edges form coherent chains of perturbation propagation which can be thought of as putative mechanisms. **Figure 8** shows a closer look of the coherent edges within the two pathways mentioned above.

For LGG, two pathways are significantly impacted with the proposed gene signature after correcting for multiple comparisons: the Ribosome pathway and the RNA transport pathway (**Figures S1**, **S2**). The reason for having only two pathways as significantly impacted could be explained by the fact that LGG is an early stage of glioma and, therefore, the differences across the given phenotypes are not reflected in the pathway level.

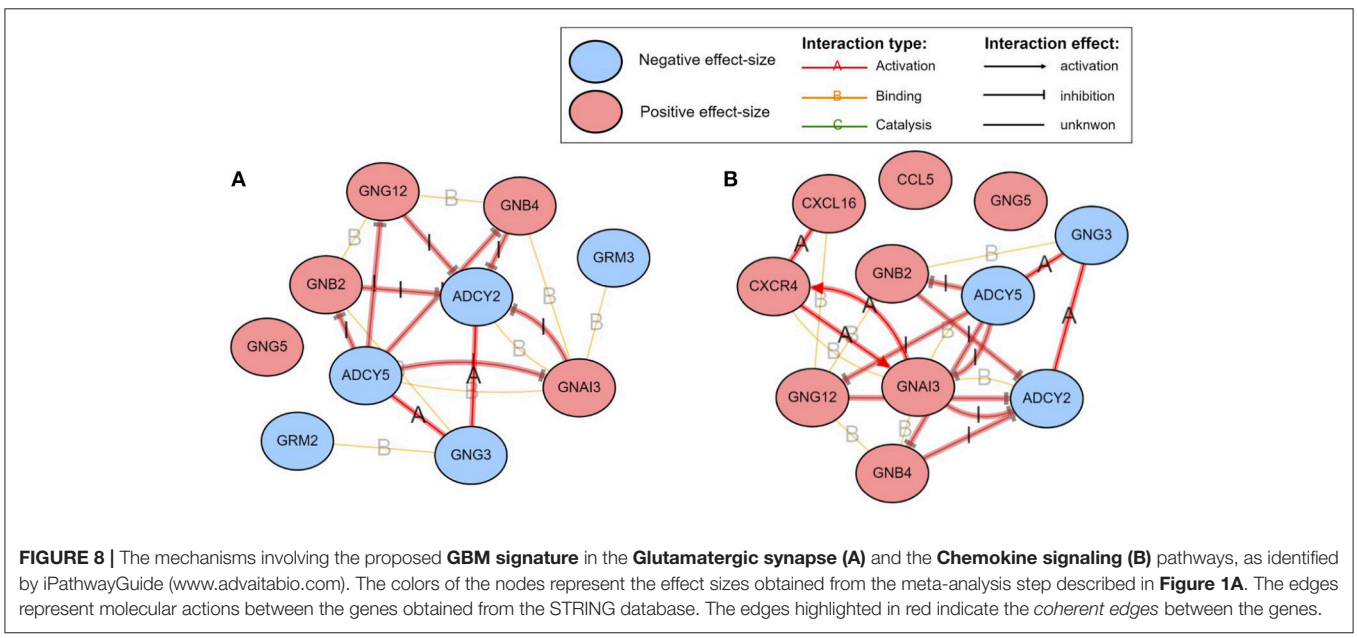


FIGURE 8 | The mechanisms involving the proposed **GBM signature** in the **Glutamatergic synapse (A)** and the **Chemokine signaling (B)** pathways, as identified by iPathwayGuide (www.advaitabio.com). The colors of the nodes represent the effect sizes obtained from the meta-analysis step described in **Figure 1A**. The edges represent molecular actions between the genes obtained from the STRING database. The edges highlighted in red indicate the *coherent edges* between the genes.

5. CONCLUSION

In an effort to identify disease-specific biomarkers that can explain the underlying biological mechanism and predict associated patients' survival, several computational approaches have been proposed over the past few years. The majority of the approaches have limited clinical applicability since they do not fully utilize the crucial information that is currently available in public repositories. In this manuscript, we propose an integrative framework that is able to identify network-based biomarkers for a given disease condition, utilizing information from three different sources: (i) multiple independent mRNA studies, (ii) multiple independent DNA methylation studies and (iii) protein-protein interactions. We demonstrate the utility of the proposed framework by constructing subnetworks related to GBM and LGG, using 17 independent mRNA and DNA methylation studies containing a total of 2,409 samples. We validate our proposed signatures on four independent gene expression datasets containing a total of 1,269 patients. The results indicate that our proposed network-based signatures are able to better predict patients' survival than other published signatures for these diseases.

AUTHOR CONTRIBUTIONS

AS and SD conceived of and designed the project. AS implemented the method in R and performed the data analysis and all computational experiments. TN, AP, and HN helped AS to perform the data analysis. AS and SD wrote the manuscript. All authors reviewed the manuscript.

FUNDING

National Institutes of Health [RO1 DK089167, STTR R42GM087013]; National Science Foundation [DBI-0965741]; and by the Robert J. Sokol M.D. Endowment in Systems Biology (to SD) DM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00159/full#supplementary-material

REFERENCES

Advaita Corporation (2019). *Pathway Analysis with iPathwayGuide*. Available online at: http://www.advaitabio.com/ipathwayguide.html

Alaimo, S., Giugno, R., Acunzo, M., Veneziano, D., Ferro, A., and Pulvirenti, A. (2016). Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* 7:54572. doi: 10.18632/oncotarget.9788

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.* 22, 2008–2017. doi: 10.1101/gr.133744.111

Ansari, S., Donato, M., Saberian, N., and Draghici, S. (2017). An approach to infer putative disease-specific mechanisms using neighboring gene networks. *Bioinformatics* 33, 1987–1994. doi: 10.1093/bioinformatics/btx097

Arimappamagan, A., Somasundaram, K., Thennarasu, K., Peddagangannagari, S., Srinivasan, H., Shailaja, B. C., et al. (2013). A fourteen gene GBM prognostic signature identifies association of immune response pathway and mesenchymal subtype with high risk group. *PLoS ONE* 8:e62042. doi: 10.1371/journal.pone.0062042

Baker, B. J., Akhtar, L. N., and Benveniste, E. N. (2009). SOCS1 and SOCS3 in the control of CNS immunity. *Trends Immunol.* 30, 392–400. doi: 10.1016/j.it.2009.07.001

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.* 33:D562–D566. doi: 10.1093/nar/gki022

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nat. Rev. Genet.* 14:333–346. doi: 10.1038/nrg3433

Burke, H. B. (2016). Predicting clinical outcomes using molecular biomarkers. *Biomark. Cancer* 8, 89–99. doi: 10.4137/BIC.S33380

Calura, E., Martini, P., Sales, G., Beltrame, L., Chiorino, G., D'Incalci, M., et al. (2014). Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Res.* 42:e96. doi: 10.1093/nar/gku354

Cancer Genome Atlas Research Network, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372, 2481–2498. doi: 10.1056/NEJMoa1402121

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028

Chang, H. Y., Nuyten, D. S., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., et al. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3738–3743. doi: 10.1073/pnas.0409462102

Chiang, J. H., Cheng, W. S., Hood, L., and Tian, Q. (2014). An epigenetic biomarker panel for glioblastoma multiforme personalized medicine through dna methylation analysis of human embryonic stem cell-like signature. *Omics* 18, 310–323. doi: 10.1089/omi.2013.0084

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Cox, D. R. (1972). Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B* 34, 187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x

Crisman, T. J., Zelaya, I., Laks, D. R., Zhao, Y., Kawaguchi, R., Gao, F., et al. (2016). Identification of an efficient gene expression panel for glioblastoma classification. *PLoS ONE* 11:e0164649. doi: 10.1371/journal.pone.0164649

Deng, L., Xiong, P., Luo, Y., Bu, X., Qian, S., and Zhong, W. (2016). Bioinformatics analysis of the molecular mechanism of diffuse intrinsic pontine glioma. *Oncol. Lett.* 12, 2524–2530. doi: 10.3892/ol.2016.5024

Diaz, D., Donato, M., Nguyen, T., and Draghici, S. (2016). "MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (New Jersey, NJ: World Scientific), 390–401.

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109. doi: 10.1016/j.tig.2005.12.005

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichiţa, C., et al. (2007). A systems biology approach for pathway level analysis. *Genome Res.* 17, 1537–1545. doi: 10.1101/gr.6202607

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *J. Psychol.* 80, 351–363. doi: 10.1080/00223980.1972.9924813

Ehtesham, M., Mapara, K. Y., Stevenson, C. B., and Thompson, R. C. (2009). CXCR4 mediates the proliferation of glioblastoma progenitor cells. *Cancer Lett.* 274, 305–312. doi: 10.1016/j.canlet.2008.09.034

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5923–5928. doi: 10.1073/pnas.0601231103

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.* 358, 1148–1159. doi: 10.1056/NEJMra072067

Etcheverry, A., Aubry, M., De Tayrac, M., Vauleon, E., Boniface, R., Guenot, F., et al. (2010). DNA methylation in glioblastoma: impact on gene expression and clinical outcome. *BMC Genomics* 11:701. doi: 10.1186/1471-2164-11-701

Fatai, A. A., and Gamieldien, J. (2018). A 35-gene signature discriminates between rapidly-and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. *BMC Cancer* 18:377. doi: 10.1186/s12885-018-4103-5

Festa, M., Caputo, M., Cipolla, C., D'Acunto, C., Rossi, A., Tecce, M., et al. (2013). The involvement of xanthohumol in the expression of annexin in human malignant glioblastoma cells. *Open Biochem. J.* 7:1. doi: 10.2174/1874091X01307010001

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., et al. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.* 64, 6503–6510. doi: 10.1158/0008-5472.CAN-04-0452

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Vol. 1. New York, NY: Springer Series in Statistics.

Garkavtsev, I., Kozin, S. V., Chernova, O., Xu, L., Winkler, F., Brown, E., et al. (2004). The candidate tumour suppressor protein ING4 regulates brain tumour growth and angiogenesis. *Nature* 428:328. doi: 10.1038/nature02329

Gevaert, O., Tibshirani, R., and Plevritis, S. K. (2015). Pancancer analysis of dna methylation-driven genes using methylmix. *Genome Biol.* 16:17. doi: 10.1186/s13059-014-0579-8

Ghiassian, S. D., Menche, J., and Barabasi, A.-L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 28, i451–i457. doi: 10.1093/bioinformatics/bts389

Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., et al. (2017). Dna methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7414–7419. doi: 10.1073/pnas.1703577114

Haynes, W. A., Vallania, F., Liu, C., Bongen, E., Tomczak, A., Andres-Terrè, M., et al. (2017). "Empowering multi-cohort gene expression analysis to increase reproducibility," in *Pacific Symposium on Biocomputing* (New Jersey, NY: World Scientific), 144–153.

Hedges, L. V., and Olkin, I. (2014). *Statistical for Meta-Analysis*. London: Academic Press.

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., de Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003. doi: 10.1056/NEJMoa 043331

Heyn, H., and Esteller, M. (2012). Dna methylation profiling in the clinic: applications and challenges. *Nat. Rev. Genet.* 13:679. doi: 10.1038/nrg3270

Hong, S. R., Jung, S. E., Lee, E. H., Shin, K. J., Yang, W. I., and Lee, H. Y. (2017). DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers. *Forensic Sci. Int. Genet.* 29, 118–125. doi: 10.1016/j.fsigen.2017.04.006

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1):S233–S240. doi: 10.1093/bioinformatics/18.suppl_1.S233

Jin, N., Wu, H., Miao, Z., Huang, Y., Hu, Y., Bi, X., et al. (2015). Network-based survival-associated module biomarker and its crosstalk with cell death genes in ovarian cancer. *Sci. Rep.* 5:11566. doi: 10.1038/srep11566

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi: 10.1038/nrg3230

Kallenberg, O. (2002). *Foundations of Modern Probability*. New York, NY: Springer-Verlag.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457–481. doi: 10.1080/01621459.1958.10501452

Kim, W., Youn, H., Lee, S., Kim, E., Kim, D., Lee, J. S., et al. (2018). RNF138-mediated ubiquitination of rps3 is required for resistance of glioblastoma cells to radiation-induced apoptosis. *Exp. Mol. Med.* 50:e434. doi: 10.1038/emm.2017.247

Kitange, G. J., Carlson, B. L., Schroeder, M. A., Grogan, P. T., Lamont, J. D., Decker, P. A., et al. (2009). Induction of MGMT expression is associated with temozolomide resistance in glioblastoma xenografts. *Neuro Oncol.* 11, 281–291. doi: 10.1215/15228517-2008-090

Klein, H. U., and Hebestreit, K. (2015). An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Brief. Bioinform.* 17, 796–807. doi: 10.1093/bib/bbv095

Komurov, K., Dursun, S., Erdin, S., and Ram, P. T. (2012). Netwalker: a contextual network analysis tool for functional genomics. *BMC Genomics* 13:282. doi: 10.1186/1471-2164-13-282

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Lan, A., Smoly, I. Y., Rapaport, G., Lindquist, S., Fraenkel, E., and Yeger-Lotem, E. (2011). ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res.* 39(Suppl 2), W424–W429. doi: 10.1093/nar/gkr359

Lee, J. H., Liu, R., Li, J., Zhang, C., Wang, Y., Cai, Q., et al. (2017). Stabilization of phosphofructokinase 1 platelet isoform by AKT promotes tumorigenesis. *Nat. Commun.* 8:949. doi: 10.1038/s41467-017-00906-9

Lee, S. Y. (2016). Temozolomide resistance in glioblastoma multiforme. *Genes Dis.* 3, 198–210. doi: 10.1016/j.gendis.2016.04.007

Lin, X. C., Zhu, Y., Chen, W. B., Lin, L. W., Chen, D. H., Huang, J.-R., et al. (2014). Integrated analysis of long non-coding RNAs and mRNA expression profiles reveals the potential role of lncRNAs in gastric cancer pathogenesis. *Int. J. Oncol.* 45, 619–628. doi: 10.3892/ijo.2014.2431

Liu, Y., Ji, Y., Li, M., Wang, M., Yi, X., Yin, C., et al. (2018). Integrated analysis of long noncoding RNA and mRNA expression profile in children with obesity by microarray analysis. *Sci. Rep.* 8:8750. doi: 10.1038/s41598-018-27113-w

Liu, Z., Yao, Z., Li, C., Lu, Y., and Gao, C. (2011). Gene expression profiling in human high-grade astrocytomas. *Comp. Funct. Genomics* 2011:245137. doi: 10.1155/2011/245137

Loskutov, Y. V., Griffin, C. L., Marinak, K. M., Bobko, A., Margaryan, N. V., Geldenhuys, W. J., et al. (2018). LPA signaling is regulated through the primary cilium: a novel target in glioblastoma. *Oncogene* 37:1457. doi: 10.1038/s41388-017-0049-3

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 27, 1290–1298. doi: 10.1093/bioinformatics/btr136

Mahajan-Thakur, S., Bien-Möller, S., Marx, S., Schroeder, H., and Rauch, B. (2017). Sphingosine 1-phosphate (S1P) signaling in glioblastoma multiforme—a systematic review. *Int J Mol Sci.* 18:2448. doi: 10.3390/ijms18112448

MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24, 1151–1161. doi: 10.1038/nbt1239

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37(Suppl 1), D619–D622. doi: 10.1093/nar/gkn863

Mitra, K., Carvunis, A. R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732. doi: 10.1038/nrg3552

Nguyen, H., Shrestha, S., Draghici, S., and Nguyen, T. (2018). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* bty1049. doi: 10.1093/bioinformatics/bty1049

Nguyen, H., Shrestha, S., Tran, D., Shafi, A., Draghici, S., and Nguyen, T. (2019). A comprehensive survey of tools and software for active subnetwork identification. *Front. Genet.* 10:155. doi: 10.3389/fgene.2019.00155

Nguyen, T., Diaz, D., and Draghici, S. (2016a). "TOMAS: a novel TOpology-aware Meta-Analysis approach applied to System biology," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Seattle, WA: ACM), 13–22.

Nguyen, T., Diaz, D., Tagett, R., and Draghici, S. (2016b). Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Nat. Sci. Rep.* 6:29251. doi: 10.1038/srep29251

Nguyen, T., Mitrea, C., and Draghici, S. (2018). Network-based approaches for pathway level analysis. *Curr. Protoc. Bioinform.* 61, 8–25. doi: 10.1002/cpbi.42

Nguyen, T., Mitrea, C., Tagett, R., and Draghici, S. (2017a). DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis. *Proc. IEEE* 105, 496–515. doi: 10.1109/JPROC.2015.2507119

Nguyen, T., Tagett, R., Diaz, D., and Draghici, S. (2017b). A novel approach for data integration and disease subtyping. *Genome Res.* 27, 2025–2039. doi: 10.1101/gr.215129.116

Nguyen, T., Tagett, R., Donato, M., Mitrea, C., and Draghici, S. (2016c). A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics* 32, 409–416. doi: 10.1093/bioinformatics/btv588

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CPG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17, 510–522. doi: 10.1016/j.ccr.2010.03.017

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Olar, A., and Sulman, E. P. (2015). Molecular markers in low-grade glioma—toward tumor reclassification. *Semin. Radiat. Oncol.* 25, 155–163. doi: 10.1016/j.semradonc.2015.02.006

Pan, Y., Smithson, L. J., Ma, Y., Hambardzumyan, D., and Gutmann, D. H. (2017). Ccl5 establishes an autocrine high-grade glioma growth regulatory circuit critical for mesenchymal glioblastoma survival. *Oncotarget* 8:32977. doi: 10.18632/oncotarget.16516

Parrella, P. (2010). Epigenetic signatures in breast cancer: clinical perspective. *Breast Care* 5, 66–73. doi: 10.1159/000309138

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13, 2363–2371. doi: 10.1101/gr.1680803

Pihur, V., Datta, S., and Datta, S. (2009). RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* 10:62. doi: 10.1186/1471-2105-10-62

Pradhan, M. P., Nagulapalli, K., and Palakal, M. J. (2012). Cliques for the identification of gene signatures for colorectal cancer across population. *BMC Syst. Biol.* 6:S17. doi: 10.1186/1752-0509-6-S3-S17

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., et al. (2013). ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res.* 41, D987–D990. doi: 10.1093/nar/gks1174

Sanz-Pamplona, R., Aragüés, R., Driouch, K., Martín, B., Oliva, B., Gil, M., et al. (2011). Expression of endoplasmic reticulum stress proteins is a candidate marker of brain metastasis in both ErbB-2+ and ErbB-2- primary breast tumors. *Am. J. Pathol.* 179, 564–579. doi: 10.1016/j.ajpath.2011.04.037

Shafi, A., Donato, M., and Drăghici, S. (2015). "A systems biology approach for the identification of significantly perturbed genes," in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* (New York, NY: ACM), 423–432. doi: 10.1145/2808719.2808763

Shafi, A., Mitrea, C., Nguyen, T., and Draghici, S. (2018). A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief. Bioinform.* 19, 737–753. doi: 10.1093/bib/bbx013

Shahbazian, D., Parsyan, A., Petroulakis, E., Hershey, J. W., and Sonenberg, N. (2010). eIF4B controls survival and proliferation and is regulated by proto-oncogenic signaling pathways. *Cell Cycle* 9, 4106–4109. doi: 10.4161/cc.9.20.13630

Shedden, K., Taylor, J. M. G., Enkemann, S. A., Tsao, M.-S., Yeatman, T. J., Gerald, W. L., et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* 14:822. doi: 10.1038/nm.1790

Shi, J., Kahle, A., Hershey, J., Honchak, B., Warneke, J., Leong, S., et al. (2006). Decreased expression of eukaryotic initiation factor 3f deregulates translation and apoptosis in tumor cells. *Oncogene* 25:4923. doi: 10.1038/sj.onc.1209495

Shi, M., Beauchamp, R. D., and Zhang, B. (2012). A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients. *PLoS ONE* 7:e41292. doi: 10.1371/journal.pone.0041292

Shi, Z., Derow, C. K., and Zhang, B. (2010). Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol.* 4:74. doi: 10.1186/1752-0509-4-74

Shukla, S., Patric, I. R., Thinagararjan, S., Srinivasan, S., Mondal, B., Hegde, A. S., et al. (2013). A DNA methylation prognostic signature of glioblastoma: identification of NPTX2-PTEN-NF-κB nexus. *Cancer Res.* 73, 6563–6573. doi: 10.1158/0008-5472.CAN-13-0298

Smith, A. A., Huang, Y. T., Eliot, M., Houseman, E. A., Marsit, C. J., Wiencke, J. K., et al. (2014). A novel approach to the discovery of survival biomarkers in glioblastoma using a joint analysis of dna methylation and gene expression. *Epigenetics* 9, 873–883. doi: 10.4161/epi.28571

Smyth, G. K. (2005). "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (New York, NY: Springer), 397–420.

Stouffer, S., Suchman, E., DeVinney, L., Star, S., and Williams, RM, J. (1949). *The American Soldier: Adjustment During Army Life*, Vol. 1. Princeton, NJ: Princeton University Press.

Sun, Y., Zhang, W., Chen, D., Lv, Y., Zheng, J., Lilljebjörn, H., et al. (2014). A glioma classification scheme based on coexpression modules of EGFR and PDGFRA. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3538–3543. doi: 10.1073/pnas.1313814111

Syed, P., Gupta, S., Choudhary, S., Pandala, N. G., Atak, A., Richharia, A., et al. (2015). Autoantibody profiling of glioma serum samples to identify biomarkers using human proteome arrays. *Sci. Rep.* 5:13895. doi: 10.1038/srep13895

Szász, A. M., Lánczky, A., Nagy, Á., Förster, S., Hark, K., Green, J. E., et al. (2016). Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* 7:49322. doi: 10.18632/oncotarget.10337

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2981–2986. doi: 10.1073/pnas.0308661100

Tarca, A. L., Drăghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., et al. (2009). A novel signaling pathway impact analysis (SPIA). *Bioinformatics* 25, 75–82. doi: 10.1093/bioinformatics/btn577

The Cancer Genome Atlas Research Network (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Turcan, S., Rohle, D., Goenka, A., Walsh, L. A., Fang, F., Yilmaz, E., et al. (2012). Idh1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* 483, 479–483. doi: 10.1038/nature10866

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., et al. (2015). DIANA-miRPath v3. 0: deciphering microRNA

function with experimental support. *Nucleic Acids Res.* 43, W460–W466. doi: 10.1093/nar/gkv403

Volinia, S., and Croce, C. M. (2013). Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 110, 7413–7417. doi: 10.1073/pnas.1304977110

Wang, J., Chen, G., Li, M., and Pan, Y. (2011). Integration of breast cancer gene signatures based on graph centrality. *BMC Syst. Biol.* 5:S10. doi: 10.1186/1752-0509-5-S3-S10

Wang, S., Jin, F., Fan, W., Liu, F., Zou, Y., Hu, X., et al. (2017). Gene expression meta-analysis in diffuse low-grade glioma and the corresponding histological subtypes. *Sci. Rep.* 7:11741. doi: 10.1038/s41598-017-12087-y

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychol. Bull.* 48:156. doi: 10.1037/h0059111

Wotschofsky, Z., Gummlich, L., Liep, J., Stephan, C., Kilic, E., Jung, K., et al. (2016). Integrated microRNA and mRNA signature associated with the transition from the locally confined to the metastasized clear cell renal cell carcinoma exemplified by miR-146-5p. *PLoS ONE* 11:e0148746. doi: 10.1371/journal.pone.0148746

Xie, Q., Wu, Q., Mack, S. C., Yang, K., Kim, L., Hubert, C. G., et al. (2015). CDC20 maintains tumor initiating cells. *Oncotarget* 6:13241. doi: 10.18632/oncotarget.3676

Yan, W., Zhang, W., You, G., Zhang, J., Han, L., Bao, Z., et al. (2012). Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma Cooperative Group. *Neuro Oncol.* 14, 1432–1440. doi: 10.1093/neuonc/nos263

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

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.

Copyright © 2019 Shafi, Nguyen, Peyvandipour, Nguyen and Draghici. 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) and the copyright owner(s) 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 Genetics

ORIGINAL RESEARCH
published: 10 May 2019
doi: 10.3389/fgene.2019.00417

Multi-Phenotype Association Decomposition: Unraveling Complex Gene-Phenotype Relationships

*Deborah Weighill[1,2], Piet Jones[1,2], Carissa Bleker[1,2], Priya Ranjan[2,3], Manesh Shah[2], Nan Zhao[3], Madhavi Martin[2], Stephen DiFazio[4], David Macaya-Sanz[4], Jeremy Schmutz[5,6], Avinash Sreedasyam[6], Timothy Tschaplinski[2], Gerald Tuskan[2] and Daniel Jacobson[1,2]**

[1] The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, United States, [2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3] Department of Plant Sciences, The University of Tennessee Institute of Agriculture, University of Tennessee, Knoxville, TN, United States, [4] Department of Biology, West Virginia University, Morgantown, WV, United States, [5] Department of Energy Joint Genome Institute, Walnut Creek, CA, United States, [6] HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council (CNR), Italy

Reviewed by:
Elena Kuzmin,
McGill University, Canada
Hugues Aschard,
School of Public Health, Harvard University, United States
Marika Kaakinen,
University of Surrey, United Kingdom

***Correspondence:**
Daniel Jacobson
jacobsonda@ornl.gov

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 19 October 2018
Accepted: 16 April 2019
Published: 10 May 2019

Citation:
Weighill D, Jones P, Bleker C, Ranjan P, Shah M, Zhao N, Martin M, DiFazio S, Macaya-Sanz D, Schmutz J, Sreedasyam A, Tschaplinski T, Tuskan G and Jacobson D (2019) Multi-Phenotype Association Decomposition: Unraveling Complex Gene-Phenotype Relationships. Front. Genet. 10:417. doi: 10.3389/fgene.2019.00417

Various patterns of multi-phenotype associations (MPAs) exist in the results of Genome Wide Association Studies (GWAS) involving different topologies of single nucleotide polymorphism (SNP)-phenotype associations. These can provide interesting information about the different impacts of a gene on closely related phenotypes or disparate phenotypes (pleiotropy). In this work we present MPA Decomposition, a new network-based approach which decomposes the results of a multi-phenotype GWAS study into three bipartite networks, which, when used together, unravel the multi-phenotype signatures of genes on a genome-wide scale. The decomposition involves the construction of a phenotype powerset space, and subsequent mapping of genes into this new space. Clustering of genes in this powerset space groups genes based on their detailed MPA signatures. We show that this method allows us to find multiple different MPA and pleiotropic signatures within individual genes and to classify and cluster genes based on these SNP-phenotype association topologies. We demonstrate the use of this approach on a GWAS analysis of a large population of 882 *Populus trichocarpa* genotypes using untargeted metabolomics phenotypes. This method should prove invaluable in the interpretation of large GWAS datasets and aid in future synthetic biology efforts designed to optimize phenotypes of interest.

Keywords: multi-phenotype associations, pleiotropy, GWAS, SNP clustering, networks, powerset space, pleiotropic signature, hypothesis generation

1. INTRODUCTION

Unraveling the complex genetic patterns underlying complex phenotypes has previously been challenging. While individual Genome-Wide Association Studies (GWAS) can provide insight into the genetic underpinnings of measured phenotypes, they typically involved associations of genetic variants with only one or a few phenotypes. The field of phenomics involves the collection of high-dimensional phenotype data of an organism, with the aim of capturing the overall, comprehensive phenotype (the "Phenome") of the organism (Houle et al., 2010). Association studies involving many measured phenotypes, for example, Phenome-Wide Association Studies (PheWAS) present many advantages, in that they allow for the complex interconnected networks between phenotypes

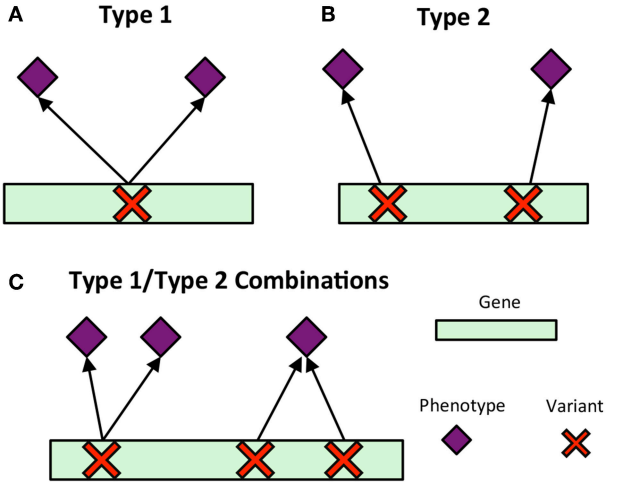


FIGURE 1 | MPA signatures. **(A)** Type 1 MPA: a gene is associated with more than one phenotype due to a single variant within the gene associating with multiple phenotypes. **(B)** Type 2 MPA: a gene is associated with more than one phenotype because of alternate SNPs within the gene having different phenotypic associations (figure created from information presented in Solovieff et al., 2013). **(C)** Complex combinations of Type 1 and Type 2 signatures.

and their genetic underpinnings to be elucidated, and also allow for the detection of pleiotropy (Pendergrass et al., 2011, 2013, 2015; Hall et al., 2014).

Pleiotropy is the phenomenon in which a gene affects multiple phenotypes (Tyler et al., 2009). One can also have a locus-centric view of pleiotropy involving a single SNP affecting multiple phenotypes (Solovieff et al., 2013). While pleiotropy used to be considered an exception to the rules of Mendelian genetics, it has since been proposed to be a common, central property inherent to biological systems (Tyler et al., 2009). Multi-phenotype associations (MPAs) can be detected in the results of Genome Wide Association Studies (GWASs) as Single Nucleotide Polymorphisms (SNPs) within genes/functional regions having multiple significant phenotype associations. This can be considered to be a pleiotropic pattern when the two phenotypes are seemingly unrelated. Two main MPA patterns exist within GWAS results. Type 1 MPAs occur when a single SNP within a functional region (such as a gene) is associated with more than one phenotype, whereas Type 2 MPAs occur when two different SNPs within a single functional region have different phenotype associations (Solovieff et al., 2013; Hackinger and Zeggini, 2017) (**Figures 1A,B**).

Multivariate analysis of the results of GWAS studies across many phenotypes have allowed for the investigation of complex relationships between genes and phenotypes, including pleiotropic relationships and the clustering of variants based on their phenotype associations. Many of these studies have involved the analysis of SNP associations with complex human disease traits. Some studies have considered pleiotropy as genes and SNPs associated with more than one phenotype, and found that pleiotropic genes tended to be longer, and that SNPs within pleiotropic genes were more likely to be exonic (Sivakumaran et al., 2011). Weighted Gene Co-expression Network Analysis (WGCNA) has been extended to cluster SNPs based on their phenotype associations using a matrix of beta coefficients, followed by hierarchical clustering of the Topological Overlap Matrix (Levine et al., 2017), and show how the resulting clusters can be used to produce polygenic scores. Gupta et al. (2011) introduced a biclustering algorithm, simultaneously clustering SNPs and phenotypes in a matrix of regression coefficients. Network-based approaches have been developed which construct bipartite networks of gene-disease phenotype associations from GWAS, and constructed network projections of this bipartite network resulting in disease similarity and gene-similarity networks (Goh and Choi, 2012). Though these studies provide a baseline of the use of multivariate and network approaches for the analysis of GWAS results, there is, to our knowledge, no method which characterizes detailed MPA signatures of genes and no method which clusters genes based on these detailed signatures. Simply clustering genes based on their phenotype associations will not capture the vast amount of combinatorial possibilities of type 1 and type 2 signatures any given gene can harbor (**Figure 1C**), especially when the multi-phenotype GWAS study involves millions of variants and hundreds of phenotypes.

Methods for multi-trait GWAS have also been developed, associating variants to groups of phenotypes (see for example Stephens, 2013; Furlotte and Eskin, 2015; Cichonska et al., 2016; Kaakinen et al., 2017a,b; Mägi et al., 2017; Porter and OReilly, 2017; Thoen et al., 2017). Mägi et al. (2017) and Kaakinen et al. (2017a) present interesting methods for identifying the association between SNPs/genes and multiple phenotypes by using the phenotypes as predictors in the modeling of the genotype. These are valuable methods for determining which phenotypes/sets of phenotypes a given gene or SNP is associated with that are more sophisticated than standard univariate GWAS approaches. These methods however do not focus on the ability to characterize and cluster genes based on the collection of topologies of SNP-phenotype associations within the gene.

We present MPA Decomposition and Signature Clustering, a network-based approach involving a constructed powerset space, in which clustering distinguishes between genes based on the detailed topology of their unique MPA signature. MPA decomposition is a post-GWAS/post-PheWAS approach with is designed to take the results of a multi-phenotype genome-wide association-type analysis (such as a standard, univariate GWAS run on several phenotypes or a multi-phenotype approach such as SCOPA (Mägi et al., 2017) and provides a framework allowing the precise mathematical representation of the architecture of variant-phenotype associations within regions (MPA/pleiotropic signatures), and thus allows these regions (such as genes) to be clustered based on these complex signatures.

2. METHODS AND MATERIALS

2.1. Overview

MPA decomposition involves the mathematical characterization of each gene's MPA signature in a network-based context. This process begins in phenotype space. In this multi-dimensional space, each axis represents a phenotype and

genes are represented as points, with points close together representing genes with similar phenotype associations and points far apart representing genes with very different phenotype associations. This phenotype space provides no information on the topology of associations within each gene. MPA decomposition maps genes to a newly constructed *powerset space*, which is constructed through clustering of SNP association vectors (**Figures 2A–E**). This clustering produces discrete sets of SNPs/overlapping sets of phenotypes called *association modules* which form the axes of powerset space, which provides the detailed structure of phenotype associations within a gene. The second stage—signature clustering—groups genes based on their detailed MPA signature (**Figure 2F**). Clustering of genes in this space results in groups of genes with identical MPA signatures. These genes grouped by MPA signatures provide a useful tool for the researcher planning genetic modification experiments, easily highlighting groups of genes with favorable signatures for modification to influence a particular phenotype.

The approach of MPA decomposition and its application are described below. MPA decomposition is a multi-step process whose results unify in a simple, matrix decomposition relationship. The multi-step process allows for the MPA signatures and signature clusters of genes to be determined from GWAS summary statistics, and is thus applicable to both newly generated genotype/phenotype data as well as published GWAS summary statistics. We apply and demonstrate this method on GWAS results from a densely genotyped *Populus trichocarpa* GWAS population involving approximately 10 million SNPs and over 400 untargetted metabolomics phenotypes measured across the population.

2.2. Metabolomics Genome-Wide Association Studies

Genotyping of 882 *P. trichocarpa* genotypes and metabolic profiling of 585 of these genotypes, followed by GWAS analysis of the 441 resulting metabolite phenotypes provided a network of associations between SNPs and metabolic phenotypes. The process for the construction of the GWAS network is described below.

2.2.1. *Populus trichocarpa* SNPs

P. trichocarpa (Tuskan et al., 2006) SNP data (DOI 10.13139/OLCF/1411410) obtained from [https://doi.ccs.ornl.gov/ui/doi/55] was derived from the whole genome resequencing of a Genome Wide Association Study (GWAS) population clonally replicated in common gardens (Tuskan et al., 2011). This dataset consists of 28,342,758 SNPs called across 882 *P. trichocarpa* genotypes. Details on the generation of this SNP dataset can be found in Weighill et al. (2018). VCFtools (Danecek et al., 2011) was used to extract the most reliable set of SNPs corresponding to the 90% tranche, resulting in a set of 10,438,861 bi-allelic SNPs.

2.2.2. Metabolomics Phenotypes

Untargetted metabolomics was conducted on *P. trichocarpa* genotypes using GC-MS. The metabolite analysis used is described in Tschaplinski et al. (2014). Briefly, samples were freeze dried for 48 h and then ground with a microWiley mill with a 20 mesh screen, with samples then twice extracted in 80% ethanol (aqueous) and the extracts combined before an aliquot was dried under nitrogen. Dried extracts were

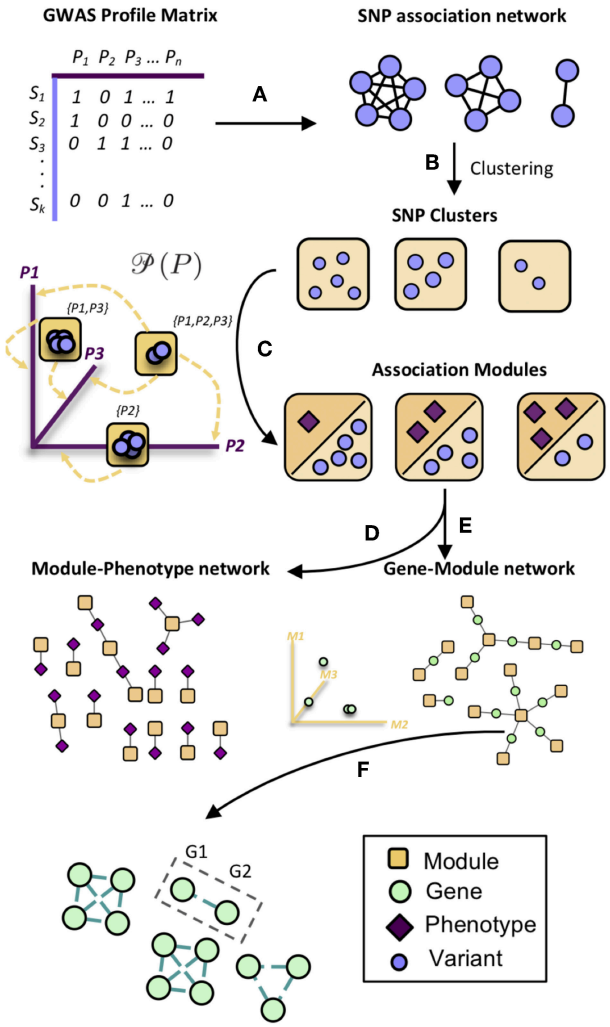


FIGURE 2 | Overview of MPA decomposition and signature clustering. **(A)** The GWAS profile matrix *M* representing SNP-phenotype associations was constructed, and the Proportional Similarity between all pairs of SNPs (rows of *M*) was calculated. **(B)** Clustering of the SNP association similarity network results in clusters of SNPs with the same phenotype associations. **(C)** Association modules are constructed as elements of the powerset of phenotypes observed in the SNP clusters. Association modules can thus be seen as non-overlapping sets of SNPs, or overlapping sets of phenotypes. These modules form the axes of powerset space. **(D)** The module-phenotype network associates the phenotypes present in each element of the powerset observed in the SNP association clusters. **(E)** The gene-module network is constructed by mapping genes to association modules if the module contains a SNP that resides within that gene. **(F)** Signature clustering is performed in *GM* (powerset) space, grouping genes with the same module associations. Clustering genes in powerset space results in groups of genes with the same pattern of MPA signatures with the same set of phenotypes. For example, a signature cluster could involve G1 and G2 containing SNPs associating with both phenotypes P1 and P2, as well as a SNP associating with only P3.

dissolved in acetonitrile followed by the addition N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane. Samples were heated for 1 h at 70°C to generate trimethylsilyl (TMS) derivatives. Samples were injected in an inert XL gas chromatograph-mass spectrometer (Agilent Technologies Inc., Santa Clara, CA, U.S.A.), fitted with an Rtx-5MS with Integra-Guard (5% diphenyl/95% dimethyl polysiloxane) capillary column (30 m by 250 μm by 0.25 μm film thickness) (Restek, Bellefonte, PA, U.S.A.). A standard quadrupole GC-MS was operated in the electron impact (70 eV) ionization mode, targeting 2.5 full-spectrum (50–650 Da) scans per second, as described previously (Tschaplinski et al., 2012). A large user-created database (>2,400 spectra) of mass spectral electron impact ionization fragmentation patterns of TMS-derivatized compounds, as well as the Wiley Registry 10th Edition with the NIST 2014 mass spectral database, were used to identify the metabolites of interest. Metabolites were quantified by extracting a key, characteristic mass-to-charge (m/z) for each known and unidentified metabolite using an automated data extraction program. Preprocessing of the resulting raw GC-MS data included alignment using XCMS (Smith et al., 2006) and normalization for amount of leaf sample analyzed, fraction of extracted sample analyzed, and internal standard recovered.

2.2.3. Outlier Analysis

We performed outlier detection on each of the metabolomic phenotypes, to account for measurement variability and technical/experimental error, using R (R Core Team, 2013). This determines which, if any, metabolite intensities that are measured over the respective genotypes (individuals), are very different from the median observed intensities for that metabolite. We applied a variant of the method discussed in Leys et al. (2013), using the median absolute deviation (MAD) from the median. Our approach differs in that it takes into account the asymmetry of the distribution of intensity values, as lower intensities are more frequent. We thus calculated the MAD for the upper and lower tails of the distribution separately. By investigating the distribution of intensities and the MAD distance from the median, for a random sample of metabolites, we determined that a MAD distance of 5 is appropriate for outlier detection, this was done using the ggplot2 package in R (Wickham, 2009). Any intensity value of a metabolite for a given genotype that was more than 5 MADs from the median was removed from the analysis. Also, to mitigate potential biases from under-represented metabolites, we excluded any metabolite that had less than 100 non-zero, non-outlier values.

2.2.4. GWAS

The EMMAX software (Kang et al., 2010) was used to statistically associate measured phenotypes with SNPs in *Populus trichocarpa*. Covariates were included to account for population structure by estimating a kinship matrix using the default parameters for Balding-Nichols method implemented in the emmax-kin program (Balding and Nichols, 1995). This was run in a parallel fashion using a customized Python script which made use of the NumPy (van der Walt et al., 2011), SciPY (http://www.scipy.org/) (Jones et al., 2001), pandas (McKinney, 2010) and mpi4py (Dalcín et al., 2005, 2008; Dalcin et al., 2011) modules. A hierarchical procedure similar to the approach described in Peterson et al. (2016), consisting of the Benjamini-Hochberg stepwise procedure (Benjamini and Hochberg, 1995) with a relaxed threshold of q1=0.1, together with the Gavrilov-Benjamini-Sarkar adaptive step-down procedure with a q2∼7.9e-06, was applied to control the false discovery rate (FDR). Associations passing the respective thresholds were considered significant associations. A total of 413 phenotypes had at least one significant SNP association, and 131,282 SNPs had at least one significant phenotype association.

2.3. MPA Decomposition

The process for MPA decomposition described below is represented visually in **Figure 2**.

2.3.1. GWAS Profile Matrix Construction

The GWAS profile matrix is the input to MPA decomposition (**Figure 2**). The GWAS profile matrix M was constructed in which each row represented a SNP that resides within a gene region, each column represented a phenotype and each entry M_{ij} was defined as:

$$M_{ij} = \begin{cases} 1 & \text{if SNP } i \text{ is associated with phenotype } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Each row of the matrix M represents the GWAS profile of a particular SNP. SNPs were mapped to their respective genes using the *P. trichocarpa* version 3 genome annotation (Tuskan et al., 2006) available on Phytozome (Goodstein et al., 2012) through the genome portal of the Department of Energy Joint Genome Institute (Grigoriev et al., 2012; Nordberg et al., 2014). A gene was considered to consist of its coding sequences as well as regulatory elements such as 5′ and 3′ UTRs.

2.3.2. Module Construction

The procedure for the construction of association modules is shown in **Figure 2**, steps A through C. The GWAS profiles of all pairs of SNPs in the GWAS profile matrix M were compared by calculating the Proportional Similarity Index between all pairs of rows of M. The Proportional Similarity Index between two vectors X and Y is defined as (Bloom, 1981):

$$PS(X, Y) = \frac{2\sum_i \min(x_i, y_i)}{\sum_i (x_i + y_i)} \quad (2)$$

where X and Y are the GWAS profiles of two SNPs (i.e., two rows of the matrix M), x_i is the ith entry in row X and y_i is the ith entry in row Y. This was performed in parallel using a customized Perl script which made use of the Parallel::MPI::Simple Perl module, developed by Alex Gough and available on The Comprehensive Perl Archive Network (CPAN) at www.cpan.org. This all-vs-all comparison results in a complete, unpruned SNP association network in which nodes represent SNPs and edges represent the similarity between the phenotype associations of SNPs.

We extracted association modules from the SNP association network as follows: First we identify SNPs that reside within genes with multiple phenotype associations (MPA genes). We extracted SNPs within MPA genes and the edges between these SNPs, and then pruned the network to only include edges between SNPs which have identical phenotype associations. This was achieved by applying a Proportional Similarity threshold of 1 (**Supplementary Texts S1, S2**). Nodes of the resulting subnetwork were then clustered into groups using MCL (Van Dongen, 2000, 2008) available from http://micans.org/mcl/. Each resulting cluster represents a group of SNPs with the same phenotype associations, i.e., a group of SNPs driven together by a particular set of phenotypes, or, an element of the powerset of phenotypes. These *modules* of phenotypes form the axes of the powerset space.

2.3.3. Module-Phenotype (MP) Matrix Construction

The *MP* matrix was constructed by mapping modules to phenotypes which drive the association between SNPs within the module (**Figure 2D**). Thus, the *MP* matrix was constructed such that each entry *ij* was defined as 1 if phenotype *j* had a significant GWAS association with all SNPs in module *i*. This could alternatively be seen as creating a network by connecting phenotype nodes to module nodes if that phenotype has a GWAS association with all SNPs in that module.

2.3.4. Gene-Module (GM) Matrix Construction

The *GM* matrix was constructed by mapping modules to genes which contained SNPs within that module (**Figure 2E**). Thus, the *GM* matrix was constructed such that each entry *ij* was defined as 1 if module *j* contained a SNP that resides within gene *i*, and zero otherwise. This can also be seen as constructing a network by connecting gene nodes to module nodes which contain SNPs that reside within that gene region.

2.3.5. Signature Clustering

Signature clustering (**Figure 2F**) was performed by calculating the similarity between all pairs of rows (genes) of the *GM* matrix using the proportional similarity metric, applying a threshold of 1, and clustering the resulting similarity network using MCL (Van Dongen, 2000, 2008).

2.4. Annotation and Functional Enrichment

P. trichocarpa gene boundaries as defined in the Ptrichocarpa_210_v3.0.gene.gff3 annotation file obtained from version 3 genome annotation (Tuskan et al., 2006) available on Phytozome was used. Functional annotations of *P. trichocarpa* genes were obtained from version 3 genome annotation (Tuskan et al., 2006) available on phytozome (Goodstein et al., 2012) through the genome portal of the Department of Energy Joint Genome Institute (Grigoriev et al., 2012; Nordberg et al., 2014).

Mapman annotations of *P. trichocarpa* were obtained by splitting the protein translations of *P. trichocarpa* genes into three sets and using the Meractor tool (Lohse et al., 2014) to assign Mapman terms to each gene. The BINGO Cytoscape plugin Maere et al. (2005) was used to determine enriched Gene Ontology (GO) terms in the set of type 1 and type 2 MPA genes.

2.5. Co-expression Network

A *P. trichocarpa* gene co-expression network was constructed as described in Weighill et al. (2018) making use of the *P. trichocarpa* (Nisqually-1) RNA-seq data derived from JGI Plant Gene Atlas project (Sreedasyam et al., unpublished), consisting of samples for various tissues (leaf, stem, root and bud tissue) and libraries generated from nitrogen source study. A list of sample descriptions was accessed from Phytozome at https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression.

3. RESULTS AND DISCUSSION

3.1. MPA Decomposition: Construction of a New Space

MPA decomposition is a multi-step process which involves the construction of a new space, allowing for the multi-phenotype signatures of genes to be easily interpreted and clustered. This method makes use of bipartite networks as data structures. Bipartite networks represent connections (edges) between two classes of objects (nodes). The results of a standard GWAS analysis were represented as a bipartite SNP-phenotype network, connecting SNP nodes to phenotype nodes between which there were significant associations. While most SNPs had only a single phenotype association, there were several SNPs which had significant associations with multiple metabolite phenotypes (**Figure 3A**). Mapping SNPs from the GWAS associations to the genes in which they reside resulted in gene-phenotype associations, which can be represented as multiple different data structures. Firstly, genes can be represented as points in multi-dimensional phenotype space, indicating their respective phenotype associations (**Figure 4**). The closer genes are to each other in phenotype space, the more shared phenotype associations they have. Alternatively, these associations can be represented as a gene-phenotype (*GP*) bipartite network, linking a gene g_i to phenotype p_k if g_i contained a SNP significantly associated with p_k (**Figure 4**). Bipartite networks are useful for the visualization and investigation of points in high dimensional space, as well as for the representation of complex relationships between multiple objects. Thus, bipartite networks were used throughout MPA decomposition as the mathematical foundation as well as a visualization tool.

GWAS associations represented as a bipartite network of SNPs connected to their associated phenotypes (**Figure 5A**) do not give any indication of MPA signatures as there is no obvious information about which SNPs belong to which genes. Thus, bipartite SNP-phenotype networks give no indication of how many phenotype associations a given gene has. GWAS associations represented as a bipartite network of genes connected to their associated phenotypes (**Figure 5B**) can give an indication as to whether or not a gene has multiple phenotype associations in that it is associated with more than one phenotype, but cannot give any indication as to the type of MPA signature (type 1 or type 2) exhibited by the gene. Mapping the SNPs in the SNP-phenotype network to the genes in which they are present results in a gene-SNP-phenotype network (**Figure 5C**). From this network, it is possible to deduce the type of MPA

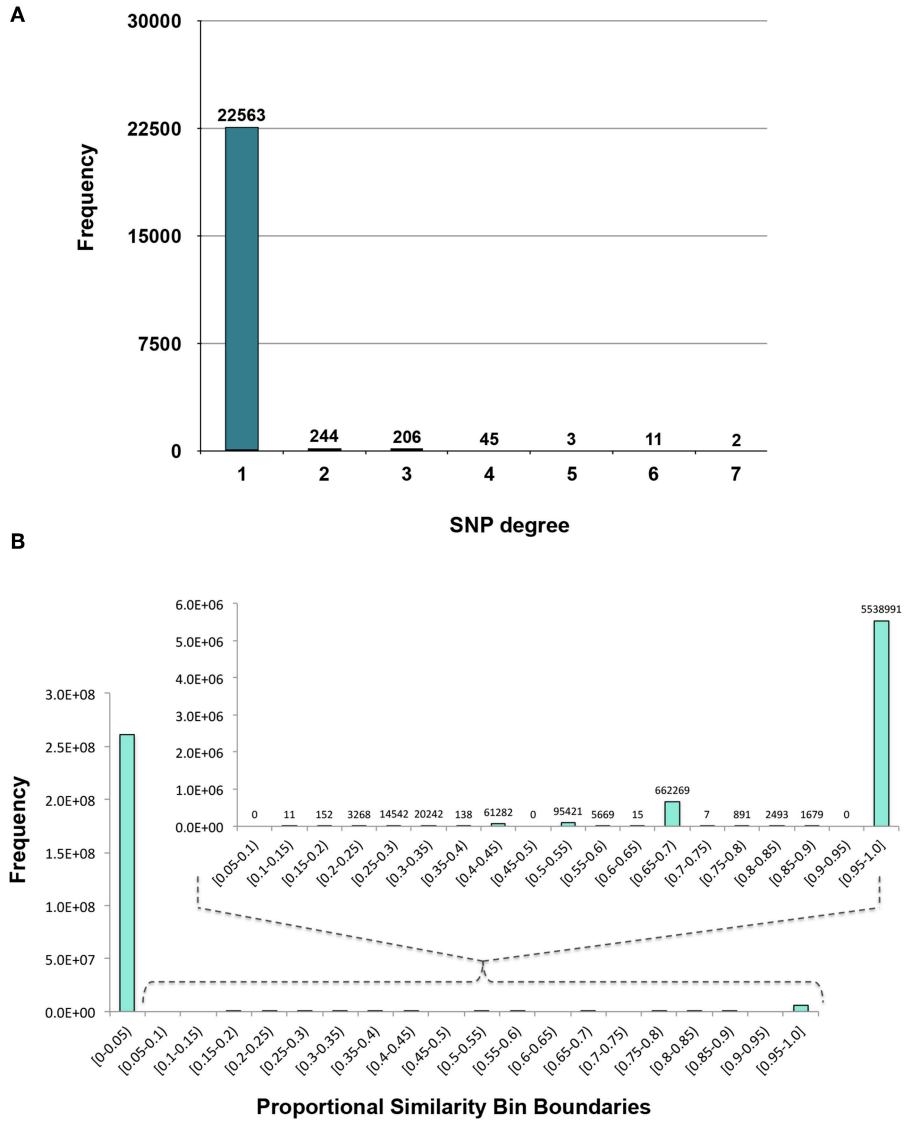


FIGURE 3 | Distributions. **(A)** Degree distribution of SNP nodes in the SNP-phenotype GWAS bipartite network. **(B)** Distribution of the Proportional Similarity edge weights in the SNP association network.

signature exhibited by a gene through some amount of visual inspection, for example, looking at the SNPs within a gene and what their associated phenotypes are. However, the structure of this network does not allow the MPA signature of a gene to be readily extracted using simple node properties such as degree. For example, one cannot simply calculate the connectivity (degree) of each gene node in **Figure 5C** in order to determine the type of MPA signature exhibited, since one can have multiple SNPs within the same gene associating with the same set of phenotypes. In addition, it is not easy to determine which genes exhibit the same MPA signatures. The process of MPA decomposition allows one to maintain the topology of SNP associations within a gene while still being able to determine the type of MPA signature using simple network measures such as degree.

The first phase of MPA decomposition involved the construction of module space, a new multi-dimensional space in which each dimension/axis represented a particular subset of phenotypes. The powerset of a set is the collection of all possible subsets of that set. Thus, we can refer to the module space as "powerset space," as each axis of the space is defined by a particular subset of phenotypes which are observed as co-associating phenotypes in the GWAS results. Modules of SNPs with the same co-associating phenotypes were identified using the Proportional Similarity metric. The distribution of Proportional Similarity values can be seen in **Figure 3B**. Of the pairs of SNPs which have non-zero Proportional Similarity values (i.e., those pairs of SNPs which shared at least one phenotype association), many had a proportional similarity value of 1.

This is explained by the degree distributions of the SNPs in the original SNP-phenotype GWAS network (**Figure 3A**). The degree distribution of a network indicates the probability (or, in this case, frequency) at which a node can be found to have a certain number of edges connected to it (Barabási and Oltvai, 2004). Therefore, the distribution in **Figure 3A** indicates that, of the SNPs which had significant phenotype associations, most of them had precisely one phenotype association. This could skew the Proportional Similarity distribution since any pairs of these "1-phenotype-hit" SNPs which are associated with the same phenotype will have a Proportional Similarity index of 1. However, it is important to keep in mind that these "1-phenotype-hit" SNPs can still contribute to MPA signatures within genes, as two "1-phenotype-hit" SNPs within the same gene that have different associations is precisely what we define as Type 2 MPA signatures.

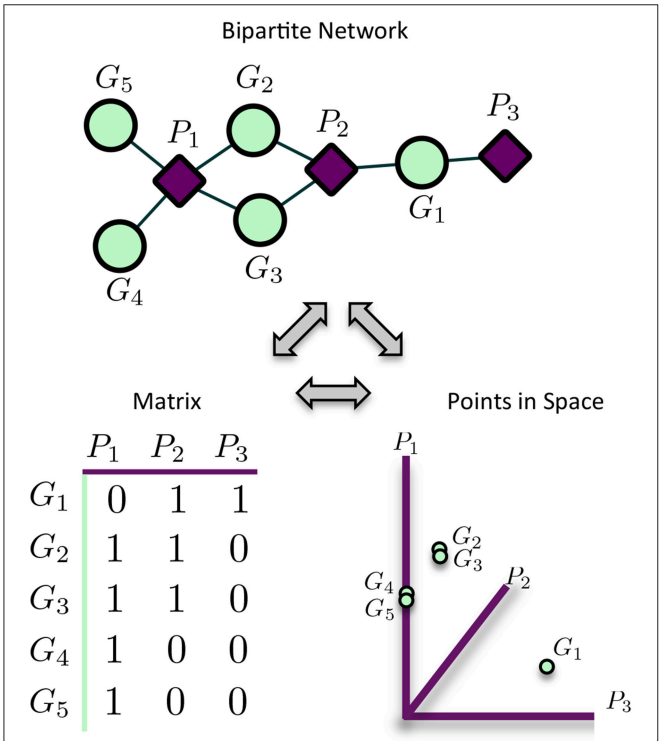


FIGURE 4 | Representation of matrices as spaces and bipartite networks. Matrices of GWAS results can easily be represented as points in high dimensional space, with rows representing points and columns representing variables/axes. Equivalently, matrices can be represented as bipartite networks, connecting row objects (genes) with column variables if the corresponding entry is non-zero. This provides a useful way to visualize high dimensional spaces as bipartite networks.

The modules form the building blocks of MPA signatures, and also conveniently collapse SNPs that are close together in genes and associate with the same set of phenotypes, and thus likely in LD. While representing non-overlapping sets of SNPs, these modules also represented overlapping sets of phenotypes. In particular, each module represented the set of phenotypes which were associated with all SNPs within the module. Thus, each module also represented an element of the powerset of phenotypes $\mathscr{P}(P)$ observed in the SNP-phenotype GWAS associations. These observed elements of the powerset were used to construct the powerset space, with each element/module representing a different dimension of this space.

These modules allowed for the construction of the gene-module (*GM*) and the module-phenotype (*MP*) matrices, which are referred to as the decomposition matrices. Represented as bipartite networks, the *MP* bipartite network defined the axes of powerset space, and the *GM* bipartite network mapped the genes into powerset space. While phenotype space provided information as to the individual phenotype associations of genes, powerset space indicated a gene's associations with sets of phenotypes at the SNP level, providing a detailed MPA signature. The mapping from phenotype space to powerset space results in a decomposition relationship between the *GP*, *GM* and *MP* matrices (**Figure 6**, **Supplementary Texts S3–S5**, **Supplementary Figure 1**). In the *GP* network (**Figure 7**), nodes

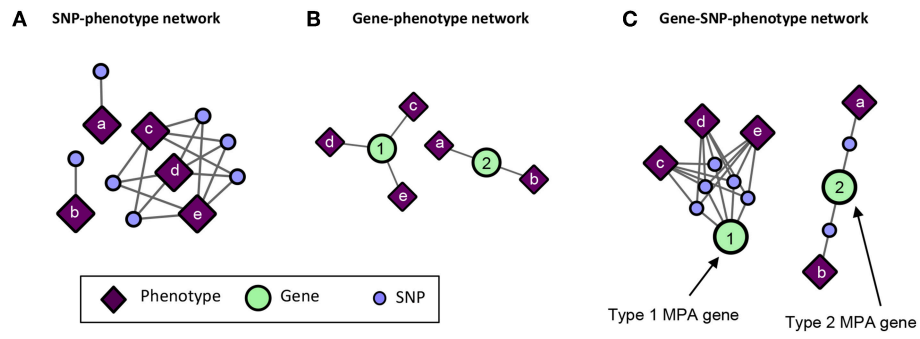


FIGURE 5 | Example of SNP-phenotype, gene-phenotype networks and gene-SNP-phenotype networks. **(A)** SNP-phenotype bipartite networks simply connect SNPs to phenotypes with which they have a significant association, and do not provide information regarding MPA signatures within genes. **(B)** Gene-phenotype networks contain connections between genes and phenotypes. An edge will be drawn between a gene and a phenotype if that gene contains a SNP associated with that phenotype. Gene-phenotype networks do not provide information as to which type of MPA signature is exhibited. **(C)** Gene-SNP-phenotype networks are SNP-phenotype networks with the SNPs connected to genes in which they reside. These networks are more complicated, and MPA signatures can be deduced from their structure through further analysis, however, the network is not in a form in which MPA signatures can be extracted easily using standard network topology measures such as degree.

represented either genes or phenotypes, and an edge was defined between gene G_i and phenotype P_j if gene G_i contained a SNP which was statistically associated with phenotype P_j in the GWAS analysis. Nodes in the *GM* network (**Figure 8**) represented either genes or modules, and an edge was defined between gene G_i and module M_j if M_j contained a SNP that resided within gene G_i. Nodes in the *MP* network (**Figure 9**) represented either association modules or phenotypes, and an edge was defined between module M_i and phenotype P_j if the correlation of SNPs within M_i is driven by phenotype P_j.

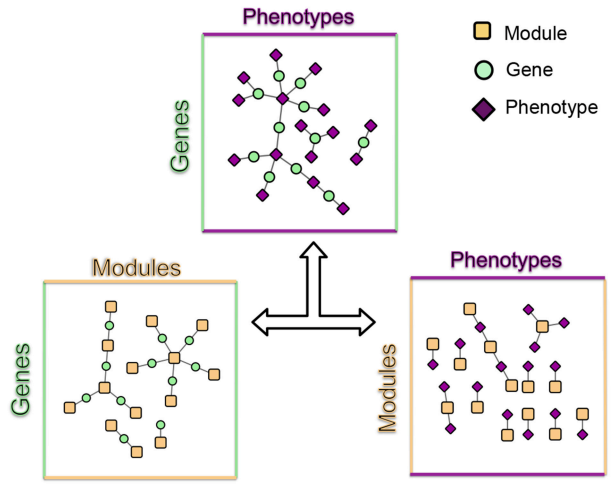


FIGURE 6 | MPA decomposition. The gene-phenotype matrix is decomposed into two matrices, a gene-module (*GM*) matrix and a module-phenotype (*MP*) matrix (**Supplementary Texts S3**, **S4**, **Supplementary Figure 1**). The *GM* matrix represents genes in powerset space. *Association modules* (elements of the powerset of phenotypes) form the basic units of MPAs and are considered latent variables. Signature clustering is performed on genes in module space (*GM* matrix).

3.2. Powerset Space Unravels Multi-Phenotype Association Signatures

The *GP* network (**Figure 7**) represents genes in phenotype space, and provides information regarding which genes are associated with which phenotypes, and can thus indicate which genes have multiple phenotype associations and are potentially pleiotropic. Of the 41,335 genes in *P. trichocarpa*, 2,964 genes had GWAS hits with more than 1 metabolite phenotype each, and are thus considered MPA genes with respect to the metabolic phenotypes.

The *GM* network (**Figure 8**) represents genes in powerset space, which in turn is defined by the *MP* network (**Figure 9**). The *GM* network unravels the MPA signatures of genes, representing their associations with sets of phenotypes. Genes that are connected to one module exhibit a Type 1 MPA signature because they contain SNPs which are associating with the same set of phenotypes, whereas genes connected to more than one module exhibit a Type 2 MPA signature because they contain SNPs which associate with different sets of phenotypes. Mapping of genes to module space thus reveals the Type 1 and Type 2 MPA patterns,

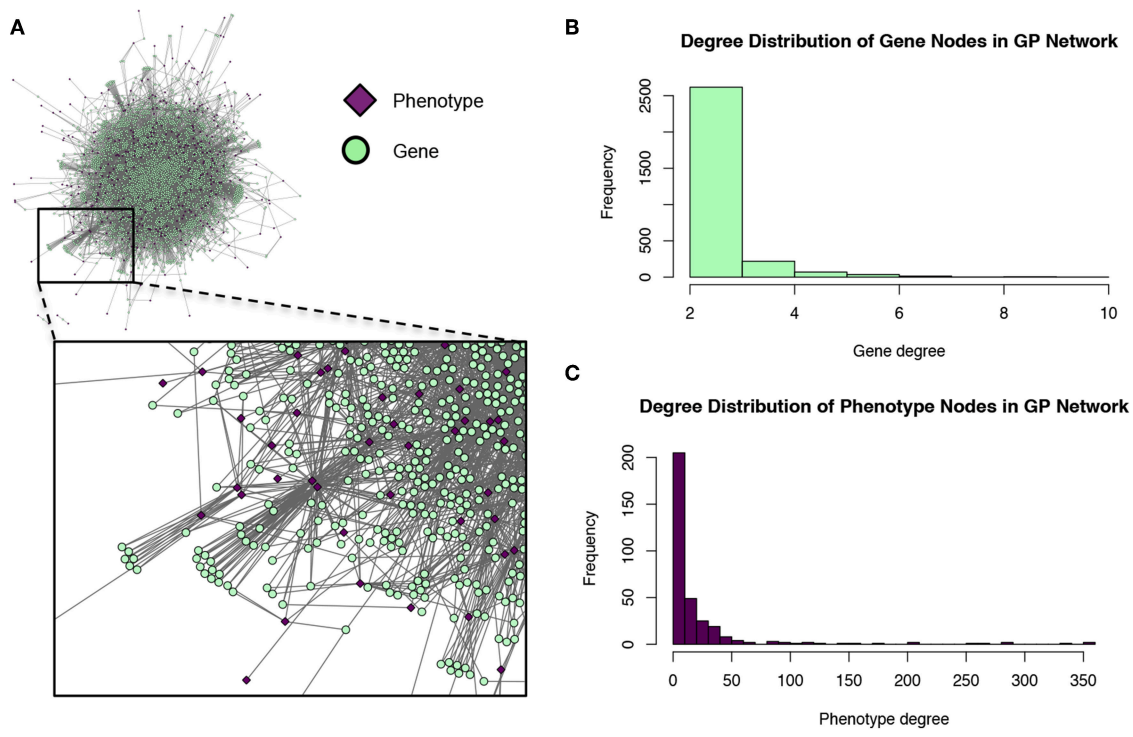


FIGURE 7 | Gene-phenotype (*GP*) network. **(A)** The *GP* network. Green nodes represent MPA genes, pink diamonds represent metabolites (phenotypes). An edge connects a gene to a phenotype if that gene contains a SNP associated with that phenotype. **(B)** Degree distribution of the gene (green) nodes in the *GP* network. **(C)** Degree distribution of the phenotype (pink) nodes in the *GP* network.

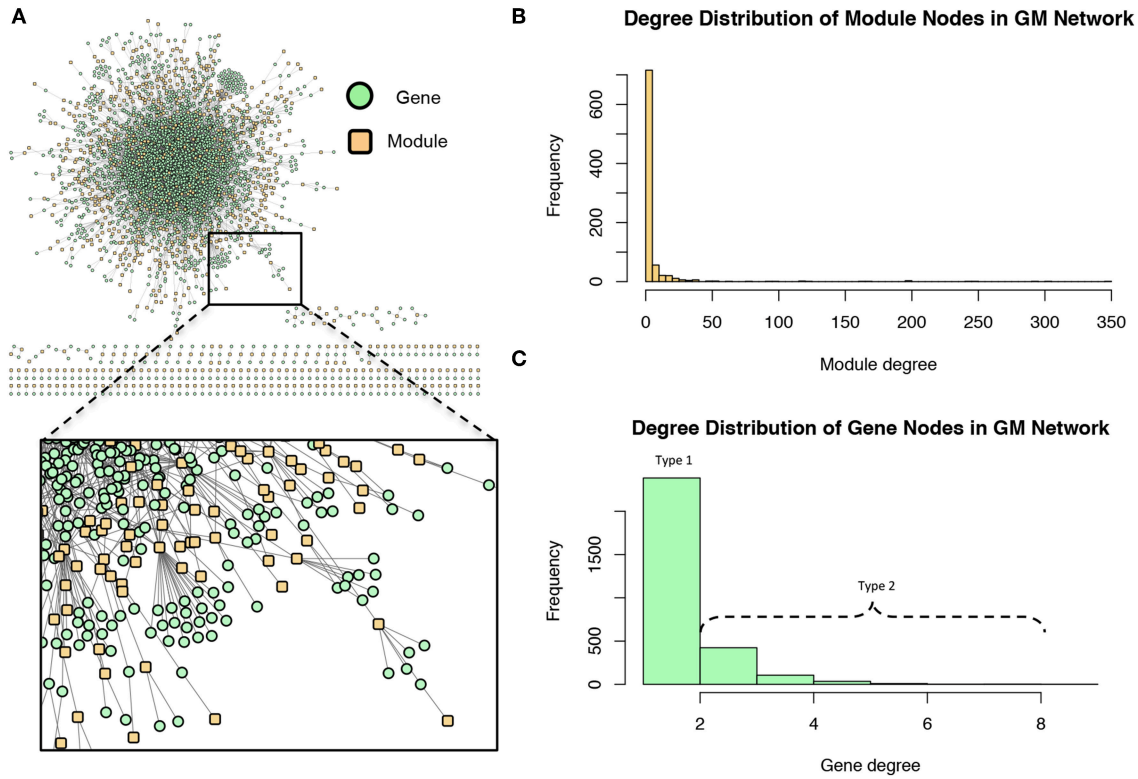


FIGURE 8 | Gene-module (*GM*) network. **(A)** The *GM* network. Green nodes represent MPA genes and yellow nodes represent association modules. A gene node is connected to a module node if the module contains a SNP which resides within that gene. **(B)** Degree distribution of the module (yellow) nodes in the *GM* network. **(C)** Degree distribution of the gene (green) nodes in the *GM* network.

as well as complex combinations of Type 1/Type 2 patterns that exist within genes (**Figure 10**). Phenotype associations of genes cannot be distinguished as Type 1 or Type 2 in phenotype space, whereas module space clearly indicates the MPA signature exhibited by a gene (**Figure 10**). Module space also goes beyond classifying genes as exhibiting Type 1 or Type 2 MPA signatures, but characterizes each unique topology of variant-phenotype associations within a gene separately. Thus, mapping of genes to module space gives information on the type of MPA signature exhibited by a gene, as well as the phenotypes involved in the signature. The high density of SNPs in this population and the rapid decay of LD allows for the high resolution of MPA signatures. **Supplementary Figure 2A** shows the variation in LD in the region including 5 kb upstream and downstream of Potri.001G419800, the type 2 MPA gene in **Figure 10F**. One can see that both associating variants in this gene are in a region of low LD. **Supplementary Figure 2B** shows a pairwise LD heatmap of 100 variants in this region including the two associating variants in Potri.001G419800. One can see that these two associating variants exist within two separate LD blocks.

The beta value derived from each SNP-phenotype association gives an indication of the effect that the SNP has on the value of the phenotype. One can look at the beta values from the GWAS analysis to see if the minor allele of a given SNP has statistically a positive or negative affect on the phenotype value. This will inform the researcher of the potential functional affect of each SNP. Overall, positive and negative beta values are present in associations in the set of type 1 MPA genes, type 2 MPA genes and single phenotype association (SPA) genes, although negative beta values are far more prevalent across all categories (**Supplementary Figure 3**) indicating that most minor alleles have negative effects on the phenotype (metabolite) values.

Of the 10,566 genes that had at least one phenotype hit, 2,964 exhibited a MPA signature by associating with more than one phenotype (**Supplementary Figure 4A**). Of those MPA genes, type 2 MPA signatures were far more abundant, with 2,468 genes exhibiting a type 2 MPA signature and 496 genes exhibiting a type 1 MPA signature (**Supplementary Table 1**, **Supplementary Figure 4B**). MPA genes represented a broad range of functions (**Figure 11**). No functional enrichment was found in the set of type 1 MPA genes. However, various GO terms were found to be enriched in the set of type 2 MPA genes, including developmental functions such as root development, shoot development, leaf development, fruit development, symbiosis, encompassing mutualism through parasitism, various regulatory functions such as RNA gene silencing function and response to stress and DNA repair (see **Supplementary Figures 5–7**, **Supplementary Table 2**, **Supplementary File 1** for complete enrichment results).

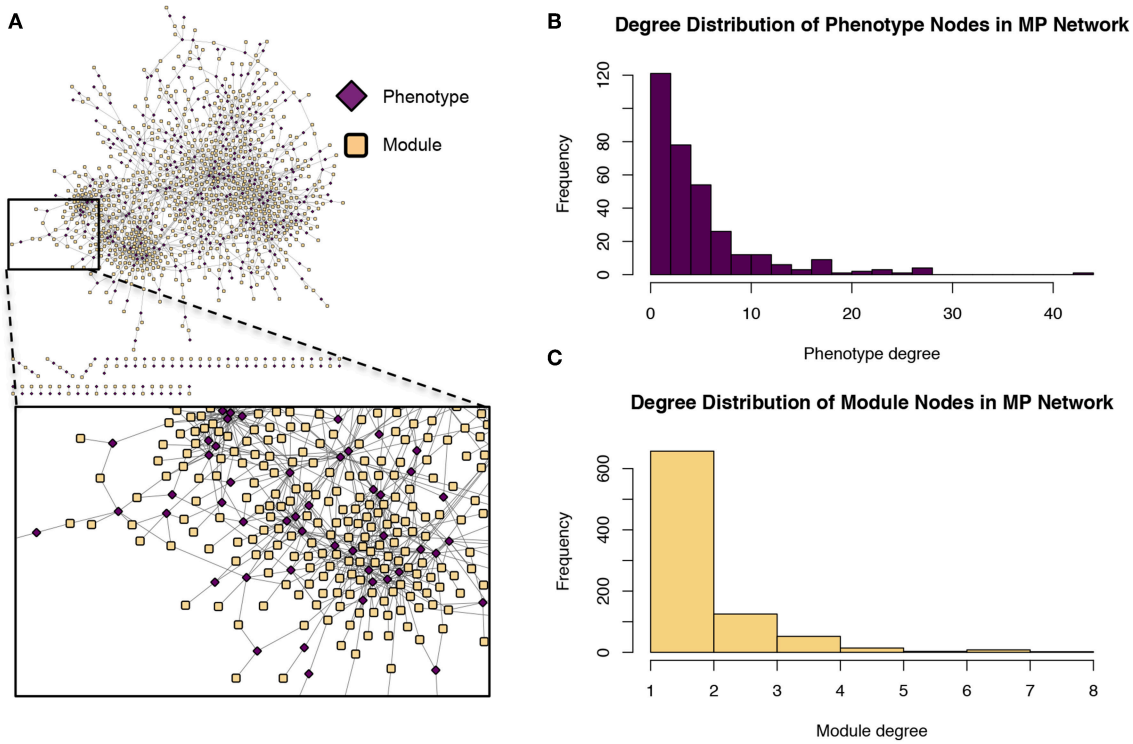


FIGURE 9 | Module-phenotype (*MP*) network. **(A)** The *MP* network. Yellow nodes represent association modules and pink nodes represent phenotypes. A module node is connected to a phenotype node if the phenotype is associated with all SNPs within the module and is thus considered a driving phenotype of the module. **(B)** Degree distribution of the phenotype (pink) nodes in the *MP* network. **(C)** Degree distribution of the module (yellow) nodes in the *MP* network.

Chaperones are classic examples of pleiotropic genes, assisting in the folding of various proteins. (Sung and Guy, 2003; Sangster et al., 2004; Gong and Golic, 2006). Querying the MPA networks for potential pleiotropic chaperones, we uncovered 14 potential chaperones based on there best *Arabidopsis* hit annotation, that contain MPA signatures (**Supplementary Table 3**), 12 of which contain type 2 MPA signatures. It is encouraging to see these classic pleiotropic genes appearing in the MPA networks, and interesting that they mostly exhibit type 2 MPA signatures.

3.3. Signature Clustering in Powerset Space

Clustering of genes in phenotype space produces groups of genes with the same overall set of phenotype associations. However, it does not provide any information as to the topology of Type 1/Type 2 associations of SNPs within the gene. Powerset space is defined by sets of phenotypes, and thus, clustering genes in this space groups genes based on the topology of Type 1/Type 2 associations of SNPs within the gene. After mapping genes to the newly constructed powerset space, genes were clustered (**Figure 2F**, *Methods and Materials*) resulting in groups of genes containing the same MPA signature. Members of a given cluster represented genes harboring identical MPA signatures. This means that genes within the same signature cluster have associations with the same modules. For example, the signature cluster driven by two modules, one involving associations with cis-3-O-caffeoyl-quinate and the other involving associations with gentisic acid-2-O-glucoside contains two genes, Potri.016G125500.v3.0 (homolog of *Arabidopsis thaliana* TRICHOME BIREFRINGENCE-LIKE 34) and Potri.012G132600.v3.0 (homolog of *Arabidopsis thaliana* AGAMOUS-like 6). These genes have associations with both cis-3-O-caffeoyl-quinate and gentisic acid-2-O-glucoside, however a given SNP within these genes is associated with either caffeoyl-quinate *or* gentisic acid-2-O-glucoside, but not both (**Figure 12**). This exemplifies what MPA decomposition and signature clustering accomplishes—the extraction of detailed multi-phenotype association signatures within genes, and the grouping of genes based on these detailed MPA signatures.

MPA signature clusters varied in size and complexity, ranging from large sets of genes having simple MPA signatures (**Supplementary Figures 8A,B**; **Supplementary Table 4**) to single gene clusters harboring very complex MPA signatures (**Supplementary Figures 8C,D**). An inverse relationship existed between the cluster size, and the number of associated phenotypes, with a minimum gene cluster size of one and a maximum gene cluster size of 42 (**Figure 13**). Complex MPA signatures are possible in this population partly because of the rapid rate with which Linkage Disequilibrium (LD) decays, dropping below 0.2 within 100 bp (**Supplementary Figure 9**).

These signature clusters are easily combined with other data types in a "lines of evidence" fashion, as introduced

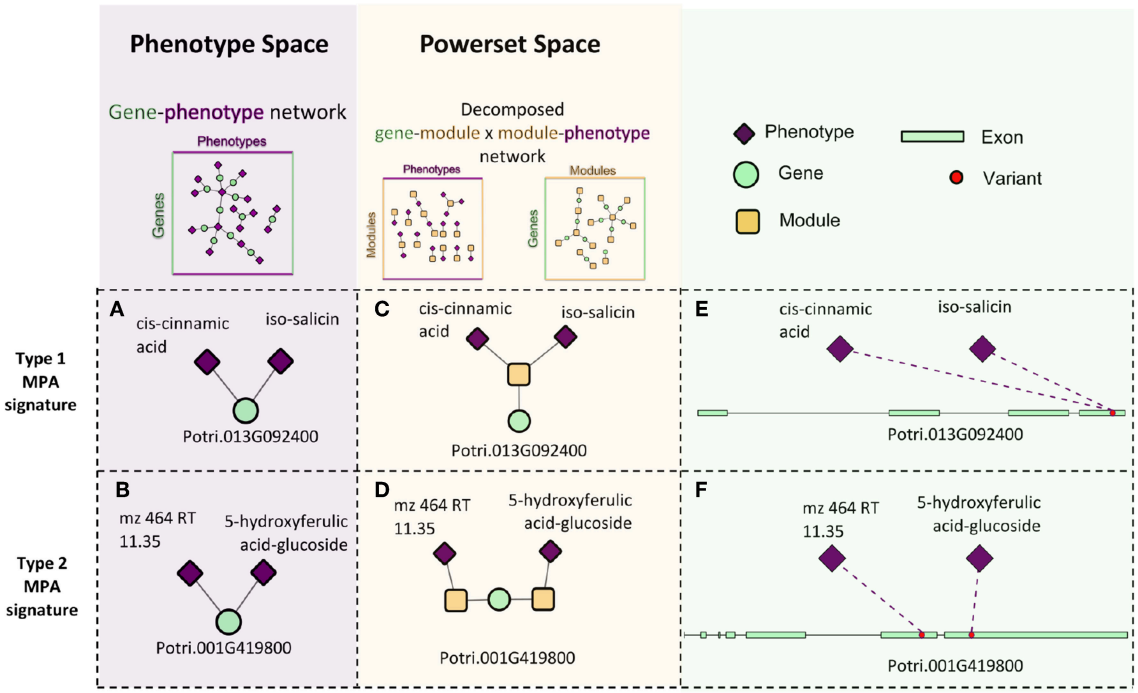


FIGURE 10 | Signature decomposition example. Two genes, Potri.013G092400 **(A)** and Potri.001G419800 **(B)** have the same surrounding network topology in the *GP* network in that they are both connected to two phenotypes. Projecting the genes into powerset space through MPA decomposition of the *GP* network indicates that they exhibit different MPA signatures in that Potri.013G092400 exhibits a type 1 MPA signature **(C)**, containing a SNP associating with two phenotypes **(E)** and Potri.001G419800 exhibits a type 2 MPA signature. **(D)** containing two SNPs, each with a different phenotype association **(F)**.

in Weighill et al. (2018). Signature clusters such as those in **Figure 12** can be merged with their neighbors in a co-expression network, providing additional insights into the functioning of these genes. Potri.016G125500 (TBL34) and Potri.012G132600 (AGL6) appeared in the same signature cluster, and are associated with many cell-wall related genes/phenotypes. TBL34 and AGL6 both associated with gentisic acid-2-O-glucoside and cis-3-O-caffeoyl-quinate, and both co-expressed with the same two transcription factors (**Figure 14**). An interesting regulatory circuit is potentially revealed, in that AGL6 potentially activates two transcription factors (positive co-expression edges) which, in turn potentially repress TBL34 (negative co-expression edges). TBL34 is also positively co-expressed with 12 genes involved in cell wall and lignin biosynthesis functions (**Figure 14**). TBL genes are known to o-acetylate xylose (Gille et al., 2011), a function which has been found to be essential for resistance to certain pathogens (Gao et al., 2017). Gentisic acid and its conjugate is a pathogen-induced signaling molecule (Bellés et al., 1999) which itself has been found to induce pathogen resistance in plants (Campos et al., 2014) and induce expression of pathogenesis-related proteins (Bellés et al., 1999). Various AGL genes are also cell-wall related in that they impact lignin content (Ferrándiz et al., 2000; Giménez et al., 2010; Cosio et al., 2017). This could be a regulatory circuit of biotic-stress-related cell wall remodeling, in which AGL6 potentially regulates xylose o-acetylation via TBL34.

3.4. Extensions to Pleiotropy

Several definitions of pleiotropy involve a gene associating with multiple, apparently disparate, unrelated phenotypes (see for example Stearns, 2010), and not all MPAs can be interpreted as pleiotropic signatures. However, if the two phenotypes are disparate enough, one can begin to hypothesize about potential pleiotropic functioning of the gene in question. In this particular study, we demonstrated our method on a collection of molecular phenotypes of metabolite concentrations. If two metabolites in a MPA exist within separate pathways, one could consider it a potentially pleiotropic interaction.

A particular example of this phenomenon found in our analysis is Potri.002G178400. This gene has a type 2 MPA association with shikimic acid and raffinose (**Supplementary Figure 10**). Based on existing knowledge found in PlantCyc on the Plant Metabolic Network (PMN) online resource (Schlapfer et al., 2017), these two metabolites are found in different pathways. Shikimic acid is involved in reactions in pathways "chlorogenic acid biosynthesis I," "chlorogenic acid biosynthesis II," "phaselate biosynthesis," "phenylpropanoid biosynthesis," "simple coumarins biosynthesis," and "chorismate biosynthesis from 3-dehydroquinate" whereas raffinose is involved in reactions in pathways "lychnose and isolychnose biosynthesis," "stellariose and mediose biosynthesis," "ajugose biosynthesis II (galactinol-independent)," "stachyose degradation," and

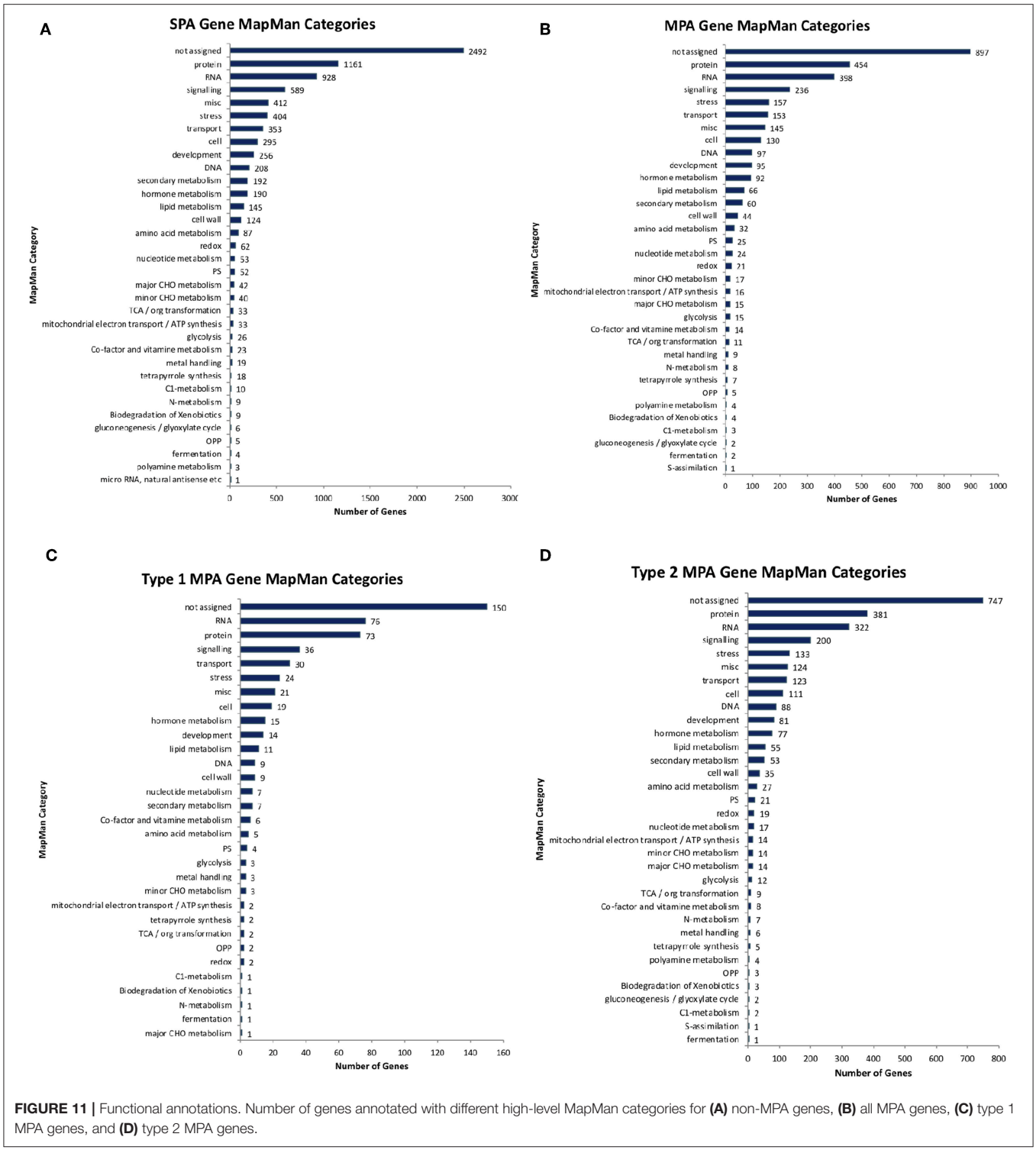


FIGURE 11 | Functional annotations. Number of genes annotated with different high-level MapMan categories for **(A)** non-MPA genes, **(B)** all MPA genes, **(C)** type 1 MPA genes, and **(D)** type 2 MPA genes.

"stachyose biosynthesis." **Supplementary File 2** contains a high resolution PDF showing the positions of raffinose (red boxes) and shikimic acid (blue box) in the *P. trichocarpa* Cellular Overview metabolic map generated on the Plant Metabolic Network online resource. Potri.002G178400 contains two Pfam domains, namely pfam01565 (FAD binding domain) and pfam04030 (D-arabinono-1,4-lactone oxidase). This is an interesting example of a potentially pleiotropic gene, which affects two different metabolic phenotypes. A possible explanation for the mechanism of this pleiotropic interaction is through competition for carbon,

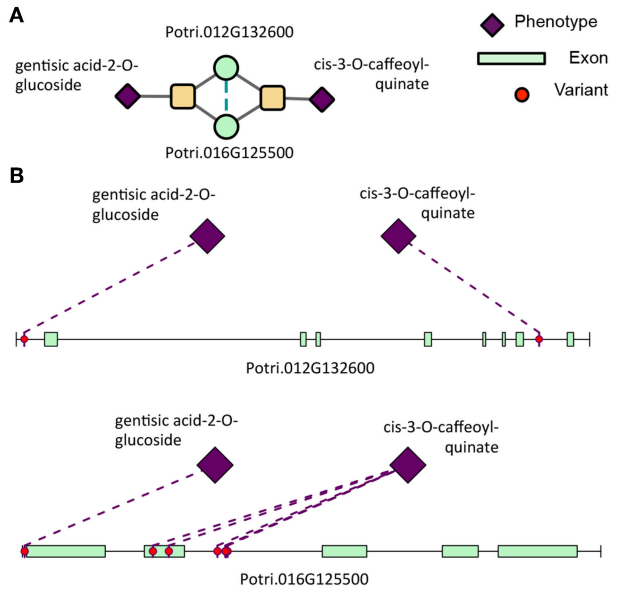


FIGURE 12 | Type 2 signature cluster. **(A)** Signature cluster defined by a Type 2 association with gentisic acid-2-O-glucoside and cis-3-O-caffeoyl-quinate. **(B)** Associating SNP positions within genes in this signature cluster. These SNP associations have negative effect sizes (beta values) on the phenotype values. See **Table 1** for gene information.

TABLE 1 | IDs, *Arabidopsis thaliana* best hits and corresponding descriptions of genes in the gentisic acid/cis-3-caffeoyl-quinate signature cluster (**Figure 12**).

Gene ID	*A. thaliana* best hit	Description
Potri.012G132600	AT2G45650	AGAMOUS-like 6
Potri.016G125500	AT2G38320	TRICHOME BIREFRINGENCE-LIKE 34

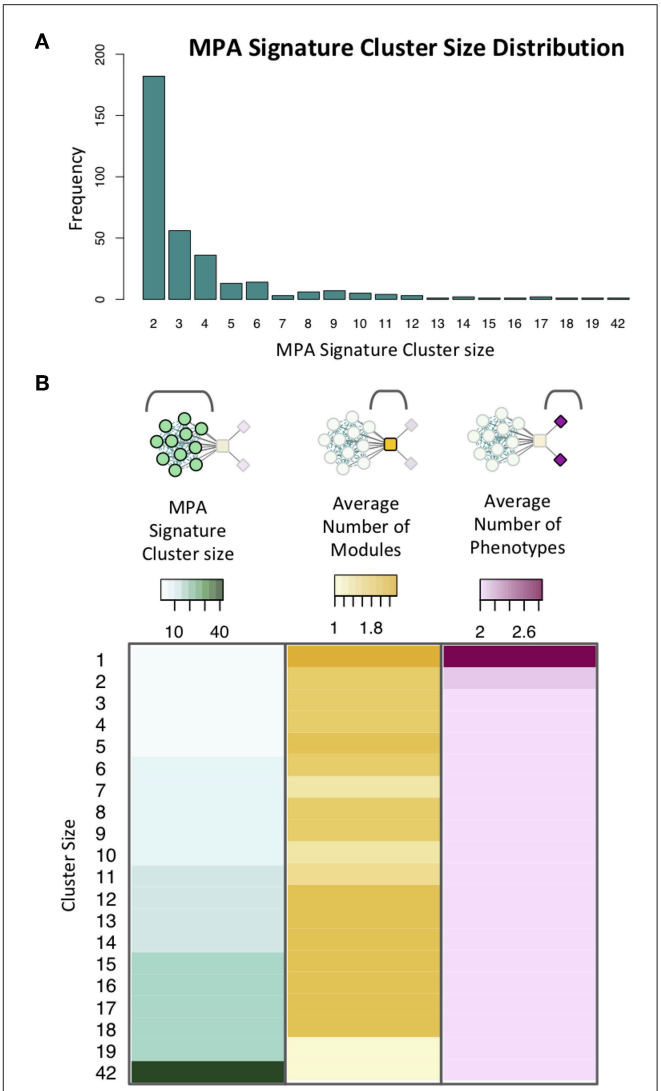


FIGURE 13 | Signature clusters in powerset space. **(A)** Cluster size distribution for signature clusters containing ≥ 2 genes. **(B)** Heatmap showing cluster size (green), average number of modules associated with genes of a given cluster size (yellow) and average number of phenotypes associated with genes in clusters of a given size (pink).

with shikimic acid committing carbon to secondary metabolism and raffinose being the product of storage for primary carbon metabolism.

It is however important to note that pleiotropic signatures can be difficult to disentangle true pleiotropic associations from other multi-phenotype associations, and should be addressed carefully. Multi-phenotype associations can be interpreted as true pleiotropy, but could also be various forms of spurious pleiotropy (see Solovieff et al., 2013 for a useful review).

3.5. Future Prospects and Implications

P. trichocarpa was an ideal species for the demonstration of the MPA decomposition for several reasons. Firstly, a large collection of 1,100 *P. trichocarpa* accessions have been clonally propagated in common gardens, resequenced and genotyped, (Tuskan et al., 2006; Slavov et al., 2012; Evans et al., 2014) providing a dense set of ∼28 million variants which are publicly available (DOI 10.13139/OLCF/1411410). Secondly, linkage disequilibrium (LD) decays very rapidly within this population of *P. trichocarpa* (**Supplementary Figure 9**). This, in combination with the dense SNP genotyping, allowed for very fine-scale MPA signatures to be resolved. Thirdly, many other different 'omics datasets exist for *P. trichocarpa* including genome scale methylation data across 10 different tissues (Vining et al., 2012) as well as a gene expression atlas are available on Phytozome (Goodstein et al., 2012). This provides extra data layers which can be integrated with the MPA networks in order to provide further interpretation and context to the GWAS associations seen in the MPA signatures, in a Lines of Evidence approach (Weighill et al., 2018). Lastly, Poplar is an important bioenergy crop (Sannigrahi et al., 2010) and is the target of extensive research. Thus, this method should be highly valuable to researchers aiming to attempt to genetically modify *P. trichocarpa* in order to impact phenotypes important to bioenergy.

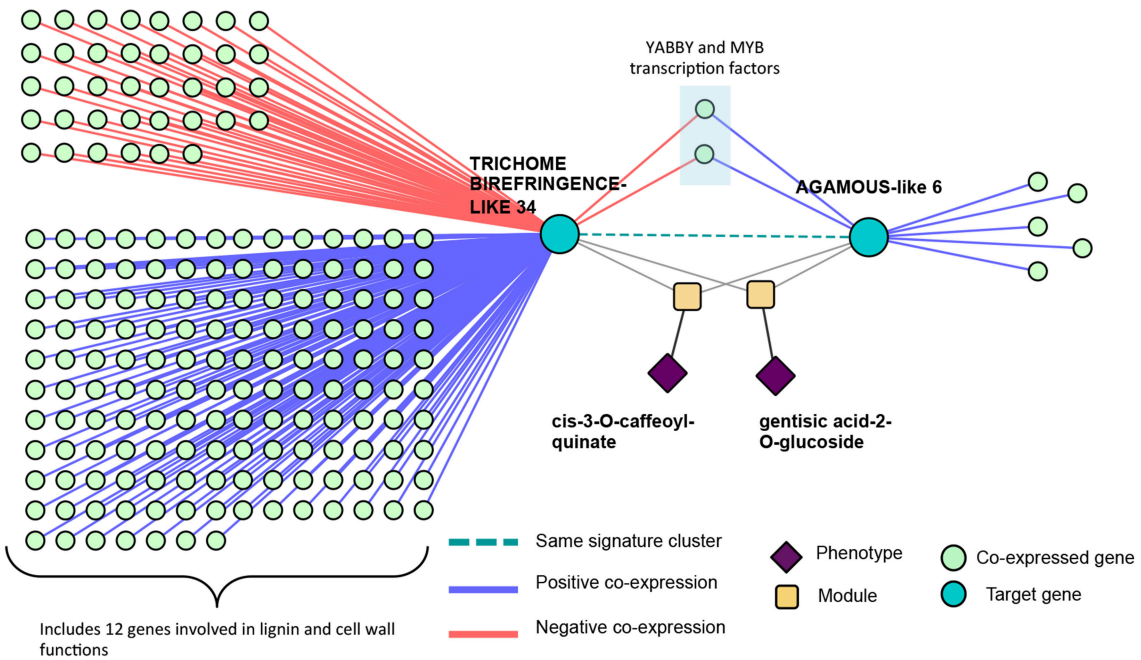


FIGURE 14 | Co-expression lines of evidence. Co-expression relationships of the signature cluster consisting of TBL34 and AGL6 from **Figure 12**.

The ease with which these MPA networks can be integrated with other network layers such as co-expression, co-methylation and SNP co-evolution networks provides a powerful strategy for furthering understanding and knowledge about the components of the system, which could aid in the annotation of genes/metabolites of previously unknown function.

Other previously published methods are able to provide information on multi-phenotype associations. The MARV (Multi-phenotype Analysis of Rare Variants) method (Kaakinen et al., 2017a) is a rare variant test that associate a gene with single or multiple phenotypes, with rare variants collapsed, so the result is gene-to-phenotype or gene-to-multi-phenotype association. This is a very valuable method to determine the potential multi-phenotype associations of a gene harboring rare variants. This method however results in a score for each gene indicating its association with a set of phenotypes, and SNP-phenotype associations within the gene are not reported. Cichonska et al. (2016) present a method of performing SNP-to-multi-phenotype and multi-SNP-multi-phenotype associations. Another method by Mägi et al. (2017) associates SNPs with multiple phenotypes through a "reverse regression" approach, using phenotypes as the predictors in the model. Both of these methods can provide a unified measure of a given variant's association with multiple phenotypes, and thus could prove to be a valuable alternative to standard univariate GWAS approaches and potentially provide an alternative, useful input set of SNP-multi-phenotype input associations to be characterized and clustered using MPA decomposition.

MPA decomposition produces signature clusters from GWAS results which can easily be merged with other data types for further interpretation. It is intended that this method will be a valuable tool in the planning of future genetic modification experiments. The resolution of the MPA signatures revealed by this method provides a useful tool to use alongside new CRISPR-based gene editing technologies to achieve high precision genome editing. This method thus provides an informed strategy for increasing the precision of future synthetic biology efforts. Researchers aiming to modify a specific gene in order to impact a particular phenotype can select genes from the signature cluster best suited to the functions they want to modify. The module decomposition also provides information as to which variants/parts of genes are associating with one phenotype or more than one phenotype, and thus can inform the researcher whether the modification of a particular location within a gene will affect more than one phenotype.

MPA decomposition will also be particularly useful in the processing and interpretation of large GWAS datasets such as eQTN studies, involving associations between millions of variants and tens of thousands of phenotypes. Future application of this method to the expanding pool of phenotypic data available will allow for the generation of comprehensive signature clusters representing the global pleiotropic potential of a given organism, and inform the planning and precision of future synthetic biology efforts to impact a wide variety and scale of phenotypes. As such, this approach should have broad impacts by developing high resolution models of MPA/pleiotropy prediction that will form the foundation of future bioengineering design efforts.

AUTHOR CONTRIBUTIONS

DJ conceived of the study and supervised the project. DW developed MPA decomposition and signature clustering,

implemented the method, generated and interpreted results and wrote the manuscript. PJ and CB performed the GWAS and outlier analysis. PR, NZ, MM, and TT performed the metabolomics. JS and AS contributed the genome sequence and transcriptome expression analysis. MS mapped gene expression atlas reads and calculated gene expression TPM values. GT led the sequencing of *Populus* genotypes. SD and DM-S performed the SNP calling and validation. DJ, GT, TT, PJ, DM-S, and SD provided editorial feedback on the manuscript.

FUNDING

Funding provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science.

This research was also supported by the Plant-Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) in the Genomic Science Program, the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science, and by the Department of Energy, Laboratory Directed Research and Development funding (7758), at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

An award of computer time was provided by the OLCF Directors Discretion program and the DOE INCITE program. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Support for the Poplar GWAS dataset was provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. The Poplar GWAS Project used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Support for DOI 10.13139/OLCF/1411410 dataset is provided by the U.S. Department of Energy, project BIF102 under Contract DE-AC05-00OR22725.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

ACKNOWLEDGMENTS

The authors would like to acknowledge the following people: Nancy Engle, David Weston, Ryan Ahg, KC Cushman, Lee Gunter, and Sara Jawdy for metabolomics sample collection. Sara Jawdy for the gene atlas experiment library preparation. Lee Gunter for preparation of the GWAS genomic samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00417/full#supplementary-material

REFERENCES

Balding, D. J., and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96, 3–12. doi: 10.1007/BF01441146

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Bellés, J. M., Garro, R., Fayos, J., Navarro, P., Primo, J., and Conejero, V. (1999). Gentisic acid as a pathogen-inducible signal, additional to salicylic acid for activation of plant defenses in tomato. *Mol. Plant Microbe Interact.* 12, 227–235. doi: 10.1094/MPMI.1999.12.3.227

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.).* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bloom, S. A. (1981). Similarity indices in community studies: potential pitfalls. *Mar. Ecol. Prog. Ser.* 5, 125–128. doi: 10.3354/meps005125

Campos, L., Granell, P., Tárraga, S., López-Gresa, P., Conejero, V., Bellés, J. M., et al. (2014). Salicylic acid and gentisic acid induce RNA silencing-related genes and plant resistance to RNA pathogens. *Plant Physiol. Biochem.* 77, 35–43. doi: 10.1016/j.plaphy.2014.01.016

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., et al. (2016). metacca: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 32, 1981–1989. doi: 10.1093/bioinformatics/btw052

Cosio, C., Ranocha, P., Francoz, E., Burlat, V., Zheng, Y., Perry, S. E., et al. (2017). The class iii peroxidase prx17 is a direct target of the mads-box transcription factor agamous-like15 (agl15) and participates in lignified tissue formation. *New Phytol.* 213, 250–263. doi: 10.1111/nph.14127

Dalcín, L., Paz, R., and Storti, M. (2005). MPI for python. *J. Parall. Distribut. Comput.* 65, 1108–1115. doi: 10.1016/j.jpdc.2005.03.010

Dalcín, L., Paz, R., Storti, M., and D'Elía, J. (2008). MPI for python: performance improvements and MPI-2 extensions. *J. Parall. Distribut. Comput.* 68, 655–662. doi: 10.1016/j.jpdc.2007.09.005

Dalcin, L. D., Paz, R. R., Kler, P. A., and Cosimo, A. (2011). Parallel distributed computing using Python. *Adv. Water Resour.* 34, 1124–1139. doi: 10.1016/j.advwatres.2011.04.013

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., et al. (2014). Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat. Genet.* 46, 1089–1096. doi: 10.1038/ng.3075

Ferrándiz, C., Liljegren, S. J., and Yanofsky, M. F. (2000). Negative regulation of the shatterproof genes by fruitfull during arabidopsis fruit development. *Science* 289, 436–438. doi: 10.1126/science.289.5478.436

Furlotte, N. A., and Eskin, E. (2015). Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model. *Genetics* 200, 59–68. doi: 10.1534/genetics.114.171447

Gao, Y., He, C., Zhang, D., Liu, X., Xu, Z., Tian, Y., et al. (2017). Two trichome birefringence-like proteins mediate xylan acetylation, which is essential for leaf blight resistance in rice. *Plant Physiol.* 173, 470–481. doi: 10.1104/pp.16.01618

Gille, S., de Souza, A., Xiong, G., Benz, M., Cheng, K., Schultink, A., et al. (2011). O-acetylation of arabidopsis hemicellulose xyloglucan requires axy4 or axy4l, proteins with a tbl and duf231 domain. *Plant Cell* 23, 4041–4053. doi: 10.1105/tpc.111.091728

Giménez, E., Pineda, B., Capel, J., Antón, M. T., Atarés, A., Pérez-Martín, F., et al. (2010). Functional analysis of the arlequin mutant corroborates the essential role of the arlequin/tagl1 gene during reproductive development of tomato. *PLoS ONE* 5:e14427. doi: 10.1371/journal.pone.0014427

Goh, K.-I., and Choi, I.-G. (2012). Exploring the human diseasome: the human disease network. *Brief. Funct. Genom.* 11, 533–542. doi: 10.1093/bfgp/els032

Gong, W. J., and Golic, K. G. (2006). Loss of hsp70 in drosophila is pleiotropic, with effects on thermotolerance, recovery from heat shock and neurodegeneration. *Genetics* 172, 275–286. doi: 10.1534/genetics.105.048793

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D. (2012). The genome portal of the department of energy joint genome institute. *Nucleic Acids Res.* 40, D26–D32. doi: 10.1093/nar/gkr947

Gupta, M., Cheung, C.-L., Hsu, Y.-H., Demissie, S., Cupples, L. A., Kiel, D. P., et al. (2011). Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome-wide associations. *J. Bone Mineral Res.* 26, 1261–1271. doi: 10.1002/jbmr.333

Hackinger, S., and Zeggini, E. (2017). Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* 7:170125. doi: 10.1098/rsob.170125

Hall, M. A., Verma, A., Brown-Gentry, K. D., Goodloe, R., Boston, J., Wilson, S., et al. (2014). Detection of pleiotropy through a phenome-wide association study (PheWAS) of epidemiologic data as part of the environmental architecture for genes linked to environment (EAGLE) study. *PLoS Genet.* 10:e1004678. doi: 10.1371/journal.pgen.1004678

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866. doi: 10.1038/nrg2897

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python*. [Online] (accessed August 16, 2016).

Kaakinen, M., Mägi, R., Fischer, K., Heikkinen, J., Järvelin, M.-R., Morris, A. P., et al. (2017a). Marv: a tool for genome-wide multi-phenotype analysis of rare variants. *BMC Bioinformatics* 18:110. doi: 10.1186/s12859-017-1530-2

Kaakinen, M., Mägi, R., Fischer, K., Heikkinen, J., Järvelin, M.-R., Morris, A. P., et al. (2017b). A rare-variant test for high-dimensional data. *Eur. J. Hum. Genet.* 25:988. doi: 10.1038/ejhg.2017.90

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Levine, M. E., Langfelder, P., and Horvath, S. (2017). "A weighted SNP correlation network method for estimating polygenic risk scores," in *Biological Networks and Pathway Analysis. Methods in Molecular Biology*, Vol. 1613, eds T. Tatarinova and Y. Nikolsky (New York, NY: Humana Press).

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49, 764–766. doi: 10.1016/j.jesp.2013.03.013

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231

Maere, S., Heymans, K., and Kuiper, M. (2005). Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551

Mägi, R., Suleimanov, Y. V., Clarke, G. M., Kaakinen, M., Fischer, K., Prokopenko, I., et al. (2017). Scopa and meta-scopa: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC Bioinformatics* 18:25. doi: 10.1186/s12859-016-1437-3

McKinney, W. (2010). "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, eds S. van der Walt and J. Millman (Austin, TX), 51–56.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., et al. (2014). The genome portal of the department of energy joint genome institute: 2014 updates. *Nucl. Acids Res.* 42, D26–D31. doi: 10.1093/nar/gkt1069

Pendergrass, S., Brown-Gentry, K., Dudek, S., Torstenson, E., Ambite, J., Avery, C., et al. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* 35, 410–422. doi: 10.1002/gepi.20589

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013). Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network. *PLoS Genet.* 9:e1003087. doi: 10.1371/journal.pgen.1003087

Pendergrass, S. A., Verma, A., Okula, A., Hall, M. A., Crawford, D. C., and Ritchie, M. D. (2015). Phenome-wide association studies: embracing complexity for discovery. *Hum. Hered.* 79, 111–123. doi: 10.1159/000381851

Peterson, C. B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies. *Genet. Epidemiol.* 40, 45–56. doi: 10.1002/gepi.21942

Porter, H. F., and OReilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait gwas methods. *Sci. Rep.* 7:38837. doi: 10.1038/srep38837

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sangster, T. A., Lindquist, S., and Queitsch, C. (2004). Under cover: causes, effects and implications of hsp90-mediated genetic capacitance. *Bioessays* 26, 348–362. doi: 10.1002/bies.20020

Sannigrahi, P., Ragauskas, A. J., and Tuskan, G. A. (2010). Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels Bioprod. Bioref.* 4, 209–226. doi: 10.1002/bbb.206

Schlapfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. *Plant Physiol.* 173, 2041–2059. doi: 10.1104/pp.16.01942

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Amer. J. Hum. Genet.* 89, 607–618. doi: 10.1016/j.ajhg.2011.10.004

Slavov, G. T., DiFazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol.* 196, 713–725. doi: 10.1111/j.1469-8137.2012.04258.x

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using monlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787. doi: 10.1021/ac051437y

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767–773. doi: 10.1534/genetics.110.122549

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* 8:e65245. doi: 10.1371/journal.pone.0065245

Sung, D. Y., and Guy, C. L. (2003). Physiological and molecular assessment of altered expression of hsc70-1 in arabidopsis. evidence for pleiotropic consequences. *Plant Physiol.* 132, 979–987. doi: 10.1104/pp.102.019398

Thoen, M. P., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P.-P., Aarts, M. G., et al. (2017). Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytol.* 213, 1346–1362. doi: 10.1111/nph.14220

Tschaplinski, T. J., Plett, J. M., Engle, N. L., Deveau, A., Cushman, K. C., Martin, M. Z., et al. (2014). *Populus trichocarpa* and *Populus deltoides* exhibit different metabolomic responses to colonization by the symbiotic fungus *Laccaria bicolor*. *Mol. Plant-Microbe Interact.* 27, 546–556. doi: 10.1094/MPMI-09-13-0286-R

Tschaplinski, T. J., Standaert, R. F., Engle, N. L., Martin, M. Z., Sangha, A. K., Parks, J. M., et al. (2012). Down-regulation of the caffeic acid *O*-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnol. Biofuels* 5:1. doi: 10.1186/1754-6834-5-71

Tuskan, G., Slavov, G., DiFazio, S., Muchero, W., Pryia, R., Schackwitz, W., et al. (2011). *Populus* resequencing: towards genome-wide association studies. *BMC Proc.* 5(Suppl. 7):I21. doi: 10.1186/1753-6561-5-S7-I21

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Tyler, A. L., Asselbergs, F. W., Williams, S. M., and Moore, J. H. (2009). Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays* 31, 220–227. doi: 10.1002/bies.200800022

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37

Van Dongen, S. (2000). *Graph Clustering by Flow Simulation* (Ph.D. thesis). University of Utrecht.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J. Mat. Anal. Appl.* 30, 121–141. doi: 10.1137/040608635

Vining, K. J., Pomraning, K. R., Wilhelm, L. J., Priest, H. D., Pellegrini, M., Mockler, T. C., et al. (2012). Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics* 13:1. doi: 10.1186/1471-2164-13-27

Weighill, D. A., Jones, P., Shah, M., Ranjan, P., Muchero, W., Schmutz, J., et al. (2018). Pleiotropic and epistatic network-based discovery: integrated networks for target gene discovery. *Front. Energy Res.* 6:30. doi: 10.3389/fenrg.2018.00030

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

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.

Copyright © 2019 Weighill, Jones, Bleker, Ranjan, Shah, Zhao, Martin, DiFazio, Macaya-Sanz, Schmutz, Sreedasyam, Tschaplinski, Tuskan and Jacobson. 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) and the copyright owner(s) 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership