



## OPEN ACCESS

## EDITED BY

Sudhindra R. Gadagkar,  
Midwestern University, United States

## REVIEWED BY

Arturo Becerra,  
National Autonomous University of  
Mexico, Mexico  
Simon Beard,  
San Sebastian University, Chile

## \*CORRESPONDENCE

Fabia Ursula Battistuzzi,  
✉ battistu@oakland.edu

RECEIVED 25 July 2025

REVISED 28 October 2025

ACCEPTED 18 November 2025

PUBLISHED 27 November 2025

## CITATION

Saravanan V, Kravetz A and Battistuzzi FU  
(2025) Higher frequency of prokaryotic low  
complexity regions in core and orthologous  
genes.

*Front. Bioinform.* 5:1673480.  
doi: 10.3389/fbinf.2025.1673480

## COPYRIGHT

© 2025 Saravanan, Kravetz and Battistuzzi.  
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.

# Higher frequency of prokaryotic low complexity regions in core and orthologous genes

Vineet Saravanan<sup>1</sup>, Alexander Kravetz<sup>2</sup> and  
Fabia Ursula Battistuzzi<sup>3,4\*</sup>

<sup>1</sup>Cranbrook Schools, Bloomfield Hills, MI, United States, <sup>2</sup>College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI, United States, <sup>3</sup>Institute for Data Science, Oakland University, Rochester, MI, United States, <sup>4</sup>Department of Biological Sciences, Oakland University, Rochester, MI, United States

Prokaryotic genome evolution is shaped by mutation, gene duplication, and horizontal gene transfer, yet the interaction of these mechanisms, particularly in relation to low complexity regions (LCRs), remains poorly understood. LCRs are known to be mutation-prone and have been proposed to promote genetic innovation. However, the interaction between LCR-mediated and paralogy-mediated genetic innovation is still unclear. To clarify the interplay between these two evolutionary forces, we analyzed the distribution of LCRs in protein-coding genes from three closely related enterobacteria (*Escherichia coli*, *Salmonella enterica*, and *Klebsiella pneumoniae*) at both species and population levels. Using pangenomic and orthology-based approaches, we categorized genes by duplication history and conservation status and assessed LCR frequencies across these groups. We found that LCRs were consistently enriched in core and orthologous genes rather than in accessory or paralogous ones. This pattern was stable across evolutionary timescales and particularly pronounced in genes involved in cell cycle control and defense. These results suggest that, contrary to prior assumptions, LCRs may serve conserved functional roles rather than acting primarily as agents of evolutionary plasticity even at population-level timescales.

## KEYWORDS

low complexity region (LCR), comparative genomics, prokaryotes, pangenome, orthologs and paralogs

## Introduction

Three major evolutionary forces that shape prokaryote genomes include the gradual accumulation of mutations, the acquisition of genes through horizontal gene transfer, and, to a lesser extent, gene duplications (Ochman et al., 2000; Ochman, 2003; Sela et al., 2016; Abby and Vincent, 2007; Lynch et al., 2023). Through these mechanisms over billions of years, prokaryotes have evolved into the large variety of population and species known today (Poole et al., 2003; Distin, 2023). While the relative frequency of these three evolutionary mechanisms has been studied before, the intersection of these three mechanisms is poorly understood (Ochman et al., 2000; Persi et al., 2023). For example, it is well known that mutational hotspots exist and that they are often located in low complexity regions (LCRs) (Haerty and Brian Golding, 2011; Teekas et al., 2024). However, the evolutionary fate of

these LCRs in duplicated or shared (i.e., horizontally transferred or evolutionarily inherited) genes is unknown.

Despite being known primarily from eukaryotes, LCRs are also present in prokaryotes, albeit in lower percentages, and are part of protein coding genes (Ntountoumi et al., 2019; Persi et al., 2023). The often repetitive structure of these regions is conducive to high mutation rates, which have led to the hypothesis of these regions being evolutionary “tuning knobs” (King et al., 1997; Kashi et al., 1997; King, 2012). The potential role of LCRs in the evolution of genomic diversity leads to two testable hypotheses: (i) the relative rate of LCRs in paralogs and ortholog should differ because of the different selective pressures acting on duplicated genes; and (ii) genes that are widely shared and conserved within and across species (core genes) should have a lower frequency of LCRs compared to those that are more sparsely distributed and more variable (accessory genes). These two hypotheses stem from a fairly straightforward scenario that correlates the high mutational rates of LCRs with gene duplication and gene sharing trends, which are driven by sequence similarity.

The availability of both population and species level data for prokaryotes offers an opportunity to test these hypotheses while also accounting for evolutionary time. The population/species boundary is expected to reflect a potential shift in evolutionary processes, from high recombination/low divergence within species to lower recombination/higher divergence among species (Touchon et al., 2009). Pangenomics has been shown to be a powerful approach to explore these trends by identifying core, conserved genes vs. accessory, more variable genes in organisms that belong to the same species or higher taxonomic units. Then, pairing this approach with orthology-detection methods that categorize genes as single or multi-copy provides the data necessary to obtain a full picture of how mutational rates in LCRs and gene duplications synergistically contribute to genomic diversity.

In this study, we apply this approach to a well-known system of three Enterobacteria species (*Escherichia coli*, *Salmonella enterica*, and *Klebsiella pneumoniae*) and their respective strains. Enterobacteria are well known species with functionally diverse strains that include both free-living and pathogenic organisms with fully sequenced genomes (Sandle, 2014). Their importance in human health, environmental microbiology, and industrial applications is well-established, thus making this work on their evolutionary plasticity significant in multiple areas of biology and beyond. By analyzing LCRs in duplicated and accessory genes, we found that the expected correlation between LCRs and high genomic diversity is not present at either the population or multi-species level. Instead, we found that LCRs are more common in orthologs and conserved (core) genes, suggesting that these regions may play a stronger functional role compared to their proposed role as “tuning knobs.”

## Methods

We obtained the complete proteomes for *Escherichia coli* str. K-12 substr. MG1655, *Salmonella enterica* subsp. *Enterica* serovar Typhimurium str. LT2, and *Klebsiella pneumoniae* subsp. *pneumoniae* HS11286 from NCBI. For the species-level analyses, we used the reference genome for each species (ASM584v2, ASM694v2,

and ASM24018v2), and for the population level analyses we randomly selected 30 genomes in each species to create three datasets of 10 strains each (Supplementary Table S1). For each species and population-level dataset, we calculated LCRs using the program SEG with default parameters (window size 12, K1 1.9, K2 2.2) (Wootton, 1994; Wootton and Scott, 1993). These parameters are commonly used to mask LCR regions during comparative analyses and, therefore, identify the most commonly identified LCRs. Additionally, these parameters identify both homo and hetero-polymers with average LCR purity (Teekas et al., 2024; Battistuzzi et al., 2016). We used ProteinOrtho with default parameters to categorize each gene as ortholog/paralog and core/accessory (Lechner et al., 2011). Genes belonging to the same ortho group and present in all 3 species (or 99% of the population genomes) were classified as core. Each gene was therefore classified in three ways: with or without LCRs, ortholog or paralog, core or accessory (Supplementary Table S2). We then calculated the frequency of LCRs as the ratio of the number of proteins with at least 1 LCR over the total number of proteins for a given category. We used chi-square tests to determine the significance of the differences observed in LCR frequencies in LCR and non-LCR containing proteins. Finally, we obtained the COG functional categories of each gene using EggNog-mapper (Cantalapiedra et al., 2021; Galperin et al., 2024) and calculated the statistical significance of over/under representation of genes with and without LCRs using a chi-square test. For each category (core, accessory, ortholog, paralog) we compared the distribution of genes with and without LCRs among different COG categories and determined its statistical significance with a chi-square test.

## Results

Genes can be categorized in two ways: (i) orthologs or paralogs, based on speciation or gene duplication histories; and (ii) core or accessory, based on their distribution among lineages. Ultimately, both of these categorizations are dependent on sequence similarity which determines how genes are related to each other within and among genomes. The presence of low complexity regions can affect their sequence similarity, as these regions are likely to have faster evolutionary rates, which results in lower sequence similarity. Thus, to determine possible correlations among LCRs, homology, and gene conservation, we performed a series of analyses at the population and species levels of *E. coli*, *S. enterica*, and *K. pneumoniae*.

First, we analyzed the three reference species for each of the species. Of the three genomes, *E. coli* and *S. enterica* are the most similar to each other as expected based on their evolutionary history (total number of proteins: 4,298 – 4,548), while *K. pneumoniae* has a larger genome (number of proteins: 5,779). Thus, relative to the total pangenome size of these three species, they contribute from 29% to 40% (Table 1).

The pangenome of these three species is composed of an approximately similar amount of core (55%) and accessory (45%) genes. However, this almost-even split changes when only protein-coding genes containing LCRs are analyzed. In this dataset, proteins containing LCRs are significantly more represented in core than in accessory genes (59% vs. 41%, chi-square test p-value <0.01) (Table 2). This result seems at odds with the high

TABLE 1 Species-level pangenome sizes.

Species	# Of proteins	Pangenome %
<i>Escherichia coli</i>	4,298	29.4
<i>Salmonella enterica</i>	4,548	31.1
<i>Klebsiella pneumoniae</i>	5,779	39.5
Total	14,625	100

Pangenome % represents that contribution of each species to the total pangenome (core+accessory genes).

TABLE 2 Percentages of pangenome core and accessory genes with or without LCRs.

Pangenomic grouping	All proteins	Proteins with LCRs	Proteins without LCRs
Pangenome	14,625	14.02% (2,051)	85.98% (12,574)
Core	54.97% (8,039)	58.8% (1,206)	54.34% (6,833)
Accessory	45.03% (6,586)	41.2% (845)	45.66% (5,741)

mutation rate of LCRs, which would be more likely to lead to lower genetic similarity and, therefore, higher probability of being in the accessory category.

This result could be explained by a long evolutionary process that filtered out the most variable LCRs, thus reducing their presence in accessory genes (Persi et al., 2023; Vishnoi et al., 2010). To test this possibility, we repeated these analyses at the population level of each species. As expected, the ratio of core-to-accessory genes is higher (>70% core, <30% accessory) reflecting the higher genetic similarity within a species. However, even in these cases, the percentages of core genes with LCRs are significantly higher (74%–80%) than those of accessory genes (20%–26%) (chi-square p-values << 0.01 for each species independently) (Table 3). These results show that higher percentage of core genes with LCRs is the same at the species and population levels, suggesting that evolutionary time does not play a role in this trend.

We repeated similar analyses with genes categorized as orthologs and paralogs. These two categories of genes are known to evolve under different selective pressures, with orthologs being more conserved and paralogs, especially recent ones, experiencing relaxed purifying selection which leads to higher evolutionary rates (Koonin, 2005; Ahrens et al., 2020). Thus, it can be hypothesized that the higher evolutionary rates of paralogs could be obtained through the accumulation of mutations in LCRs. Interestingly, we found that this was not the case. Instead, we found a higher probability of genes with LCRs in the ortholog category than in paralogs (chi-square test p-values << 0.01), again supporting the hypothesis that genes with LCRs are more likely to be conserved than those without LCRs. It is also possible that the relative lengths of proteins and LCRs in orthologs and paralogs is different, leading to a different evolutionary behavior of LCRs in these two categories. To test this

possibility, we performed Mann Whitney U tests on the LCR lengths of orthologs vs. paralogs and found a significant difference (p-value 0.006). However, upon closer examination, we found that this result is driven purely by 23 LCRs in a single gene (StfR) which is a putative prophage side tail fiber (NP\_415890.2). Upon elimination of this outlier, the p-value rises to 0.82 showing that there is no statistically significant difference in LCR length between orthologs and paralogs. Similarly, the protein length between orthologs and paralogs is not statistically significantly different (p-value 0.06 and 0.28 without StfR).

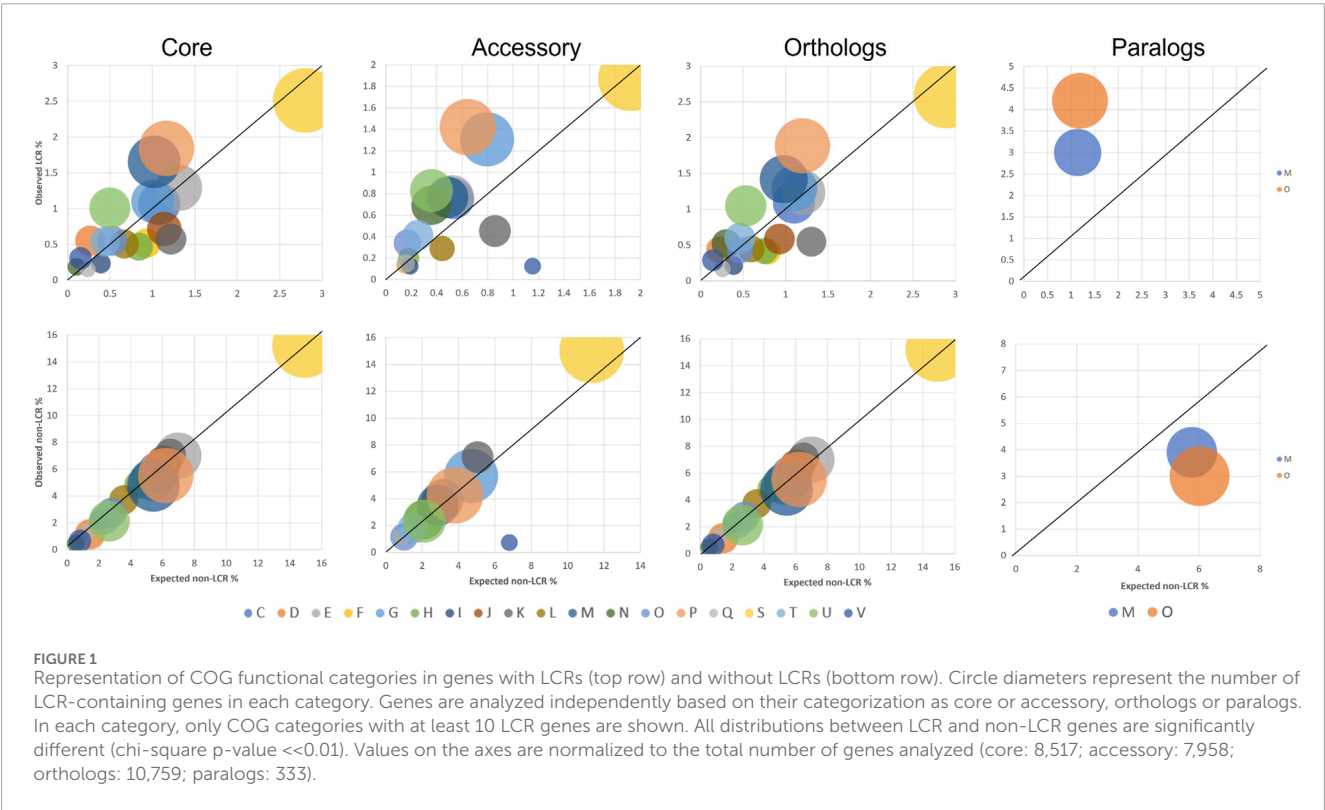
From a functional perspective, core/accessory and orthologous/paralogous genes have different distributions among COG categories. To identify potential correlations between COG functions and LCR distributions, we investigated the patterns in functional category for genes with and without LCRs. We performed this analysis at the species level as gene functions within species are largely conserved (Santoni and Romano-Spica, 2009). The overall distribution of genes with and without LCRs is significantly different across COG functional categories in each group of core, accessory, orthologs, and paralogs (chi-square p-values << 0.01 in each group). Some categories show clear trends of enrichment or depletion in LCR vs. non-LCR genes: enriched in LCRs are D (Cell cycle control, cell division, chromosome partitioning), N (Cell motility), O (Post-translational modification, protein turnover, chaperon functions), P (Inorganic ion transport and metabolism), T (Signal transduction mechanisms), and U (Intracellular trafficking, secretion, and vesicular transport); depleted are F (Nucleotide metabolism and transport), I (Lipid metabolism), J (Translation, ribosomal structure and biogenesis), K (Transcription), L (Replication, recombination, and repair), Q (Secondary structure), and also the unknown function (S) (Figure 1). These functional associations are similar to those previously found, in particular with LCRs being involved in transport, signal transduction, and trafficking, genes that are evolutionarily dynamic (Wright and Jane Dyson, 2015; Dyson and Wright, 2005; Puigbò et al., 2014). However, within LCR-containing genes specifically, two functional categories (D and V (Defense mechanisms)) are enriched in core and orthologs, and four are enriched only in accessory and/or paralogs (C (Energy production and conversion), E (Amino acid metabolism and transport), H (Coenzyme metabolism and transport), and O (Post-translational modification, protein turnover, chaperon functions)) (Figure 1).

## Discussion

The accumulation of mutations and the acquisition/loss/duplication of genes among genomes are primary mechanisms that lead to genomic diversity. While comparative genomics has been traditionally based on sequence similarity, the concept of a pangenome expanded this approach to include gene presence/absence as a measure of similarity among genomes. Nonetheless, pangenomes still rely on sequence similarity measures to cluster genes and, therefore, provide a way to connect these two mechanisms: from genes that are highly similar and shared across multiple lineages (core) to genes that are so different that they are not clustered together, thus forming groups of accessory genes (i.e., missing in some of the genomes). Additionally, core

TABLE 3 Population pangenome size, core, and accessory gene percentages. Data for genes with and without low complexity regions (LCRs) is shown. Ranges for three random population datasets for each species are shown in parenthesis.

Pangenomic grouping	<i>E. coli</i>		<i>S. enterica</i>		<i>K. pneumoniae</i>	
Pangenome size	47,689 (47,000–48190)		43,322 (42,279–45392)		49,264 (48,632–49781)	
Core	69.7% (66%–73.5%)		73.5% (72.7%–75%)		76% (72.6%–78.7%)	
Accessory	30.3% (26.5%–34%)		27% (25%–27.3%)		24 (21.3%–27.4%)	
	LCRs	Non-LCRs	LCRs	Non-LCRs	LCRs	Non-LCRs
Core	74.3% (69.5%–78.4%)	69.1% (65.5%–72.9%)	79.6% (78.2%–81.7%)	72.6% (71.6%–74.4%)	80.2 (77.3%–82.9%)	75.12% (71.7%–77.9%)
Accessory	25.7% (21.6–30.5)	30.9% (27.1%–34.5%)	20.4% (18.3%–21.8%)	27.4% (25.6%–28.4%)	19.8% (17.1%–22.7%)	24.8% (22.1%–28.3%)



and accessory genes can also be either single or multi-copy, which also can affect sequence similarity by altering selective pressures on different copies (Innan and Kondrashov, 2010; Kondrashov, 2012; Qian and Zhang, 2014; Vishnoi et al., 2010). The manifestation of all these different evolutionary mechanisms is a change in sequence similarity. However, this change can be small, especially when observing population level data. Thus, it is convenient to focus on regions in a genome with high mutation rates, such as low complexity regions. In addition to being highly variable, these regions have also been proposed as key players in evolutionary innovations (“tuning knobs”), which would suggest a more widespread presence in specialized (i.e., non-conserved) genes (Kashi et al., 1997; King, 2012).

To test this prediction, we analyzed population and species level data for three closely related bacteria, *E. coli*, *S. enterica*, and *K.*

*pneumoniae*, to determine the patterns of association between LCRs and sequence conservation. As a proxy for sequence conservation, we used orthology/paralogy and core/accessory categorizations. Contrary to the expectation of an association between LCR and low sequence conservation (accessory and paralogous genes), our results show that both at the species and population level, LCRs are primarily present in core and orthologous genes, which are most likely to be more conserved than accessory and paralogs (Koonin, 2005; Jordan et al., 2002). The presence of the same trend at different timescales (populations and species) suggests that this is a functionality-driven pattern rather than the result of time leading to the filtering of highly divergent sequences. Thus, this result suggests one primary conclusion: LCRs are important elements in conserved proteins. The association of LCRs with orthologs supports a recent a study that proposed LCRs and gene duplication as two sequential

processes leading to genomic diversity (Persi et al., 2023). However, our results expand this model by showing a preferential association of LCRs with core and orthologous genes even at the population level, thus suggesting that the separation of the LCR and gene duplication evolutionary mechanisms happens very early in the history of a species.

Interestingly, the core and orthologous genes with high LCR frequency are primarily in the cell cycle and defense functional categories, while those in the accessory and paralogous genes are in metabolism, post-translational modifications, and energy production functions. This result could be explained in different ways. For example, it is possible that once a gene has obtained its function, the LCRs embedded in the gene lose their high variability in favor of maintaining the function of the gene. However, this interpretation disagrees with the known lower conservation of genes especially in the defense category (Makarova et al., 2011; Makarova et al., 2013). Thus, an alternative interpretation is that LCRs in these genes actually maintain their high variability rate and provide necessary diversity to genes that require constant adaptation but that are otherwise conserved (Kashi et al., 1997; King, 2012; Velasco and Hernández-Morales, 2013; King and Kashi, 2007). The two scenarios are not mutually exclusive and are likely to explain different subsets of LCRs. Irrespective of these scenarios, the similarity of our results at both population and species level suggests that the functional role of the LCRs is gained relatively quickly after their evolution and is then maintained over time. Future analyses that quantify the conservation of genes within and outside of their LCRs will allow to determine the specific functional role played by these regions.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/>, ASM584v2, ASM694v2, and ASM24018v2.

## Author contributions

VS: Methodology, Formal Analysis, Writing – original draft. AK: Formal Analysis, Writing – original draft, Methodology. FB: Conceptualization, Funding acquisition, Methodology, Supervision, Visualization, Writing – review and editing.

## Funding

The authors declare that financial support was received for the research and/or publication of this article. This work was partially

supported by a National Science Foundation National Research Traineeship program award (#2299041) to F.U.B.

## Acknowledgements

We thank two reviewers for their insights which have strengthened this manuscript during revisions.

## 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbinf.2025.1673480/full#supplementary-material>

## References

- Abby, S., and Vincent, D. (2007). Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.* 15 (3), 135–141. doi:10.1016/j.tim.2007.01.007
- Ahrens, J. B., Teufel, A. I., and Siltberg-Liberles, J. (2020). A phylogenetic rate parameter indicates different sequence divergence patterns in orthologs and paralogs. *J. Mol. Evol.* 88 (10), 720–730. doi:10.1007/s00239-020-09969-7
- Battistuzzi, F. U., Schneider, K., Spencer, M. K., Fisher, D., Chaudhry, S., and Escalante, A. A. (2016). Profiles of low complexity regions in apicomplexa. *BMC Evol. Biol.* 16, 47. doi:10.1186/s12862-016-0625-0
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-Mapper v2: functional annotation, orthology assignments, and



- domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38 (12), 5825–5829. doi:10.1093/molbev/msab293
- Distin, M. R. (2023). “Genetic evolvability: using a restricted pluralism to tidy up the evolvability concept,” in *Evolutionary biology: contemporary and historical reflections upon core theory*. Editors T. E. Dickins, and J. A. Benjamin (Springer International Publishing). doi:10.1007/978-3-031-22028-9\_38
- Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6 (3), 197–208. doi:10.1038/nrm1589
- Galperin, M. Y., Alvarez, R. V., Karamycheva, S., Makarova, K., Wolf, Y., Koonin, S., et al. (2024). COG database update 2024. *Nucleic Acids Res.* 4, gkae983. doi:10.1093/nar/gkae983
- Haerty, W., and Brian Golding, G. (2011). Increased polymorphism near low-complexity sequences across the genomes of *Plasmodium falciparum* isolates. *Genome Biol. Evol.* 3, 539–550. doi:10.1093/gbe/evr045
- Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11 (2), 97–108. doi:10.1038/nrg2689
- Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12 (6), 962–968. doi:10.1101/gr.87702
- Kashi, Y., King, D., and Morris, S. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 13 (2), 74–78. doi:10.1016/S0168-9525(97)01008-1
- King, D. G. (2012). “Evolution of simple sequence repeats as mutable sites,” in *Tandem repeat polymorphisms: genetic plasticity, neural diversity and disease*. Editor A. J. Hannan (Springer). doi:10.1007/978-1-4614-5434-2\_2
- King, D. G., and Kashi, Y. (2007). Mutability and evolvability: indirect selection for mutability. *Heredity* 99 (2), 123–124. doi:10.1038/sj.hdy.6800998
- King, D. G., Morris, S., and Yechezkel, K. (1997). Evolutionary tuning knobs. *Endeavour* 21 (1), 36–40. doi:10.1016/S0160-9327(97)01005-3
- Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.* 279 (1749), 5048–5057. doi:10.1098/rspb.2012.1108
- Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39 (1), 309–338. doi:10.1146/annurev.genet.39.073003.114725
- Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (Co-)Orthologs in large-scale analysis. *BMC Bioinforma.* 12, 124. doi:10.1186/1471-2105-12-124
- Lynch, M., Ali, F., Lin, T., Wang, Y., Ni, J., and Long, H. (2023). The divergence of mutation rates and spectra across the tree of life. *EMBO Rep.* 24 (10), e57561. doi:10.15252/embr.202357561
- Makarova, K. S., Wolf, Y. I., Snir, S., and Koonin, E. V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* 193 (21), 6039–6056. doi:10.1128/jb.05535-11
- Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* 41 (8), 4360–4377. doi:10.1093/nar/gkt157
- Ntountoumi, C., Vlastaridis, P., Mossialos, D., Stathopoulos, C., Iliopoulos, I., Promponas, V., et al. (2019). Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic Acids Res.* 47 (19), 9998–10009. doi:10.1093/nar/gkz730
- Ochman, H. (2003). Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.* 20 (12), 2091–2096. doi:10.1093/molbev/msg229
- 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
- Persi, E., Wolf, Y. I., Karamycheva, S., Makarova, K. S., and Koonin, E. V. (2023). Compensatory relationship between low-complexity regions and gene paralogy in the evolution of prokaryotes. *Proc. Natl. Acad. Sci.* 120 (16), e2300154120. doi:10.1073/pnas.2300154120
- Poole, A. M., Phillips, M. J., and Penny, D. (2003). Prokaryote and eukaryote evolvability. *Biosystems* 69 (2), 163–185. doi:10.1016/S0303-2647(02)00131-4
- Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I., and Koonin, E. V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12 (1), 66. doi:10.1186/s12915-014-0066-4
- Qian, W., and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Res.* 24 (8), 1356–1362. doi:10.1101/gr.172098.114
- Sandle, T. (2014). “Biochemical and modern identification techniques | enterobacteriaceae, coliforms, and *Escherichia Coli*,” in *Encyclopedia of food microbiology*. Editors C. A. Batt, and M. Lou Tortorello Second Edition (Academic Press). doi:10.1016/B978-0-12-384730-0.00037-9
- Santoni, D., and Romano-Spica, V. (2009). Comparative genomic analysis by microbial COGs self-attraction rate. *J. Theor. Biol.* 258 (4), 513–520. doi:10.1016/j.jtbi.2009.01.035
- Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci.* 113 (41), 11399–11407. doi:10.1073/pnas.1614083113
- Teekas, L., Sharma, S., and Vijay, N. (2024). Terminal regions of a protein are a hotspot for low complexity regions and selection. *Open Biol.* 14 (6), 230439. doi:10.1098/rsob.230439
- Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., et al. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *Plos Genet.* 5 (1), e1000344. doi:10.1371/journal.pgen.1000344
- Velasco, M., Hernández-Morales, R., Delaye, L., Jiménez-Corona, M. E., Ponce-de-Leon, S., et al. (2013). Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 Gp120 protein. *J. Theor. Biol.* 338 (December): 80–86. doi:10.1016/j.jtbi.2013.08.039
- Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannehalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Res.* 20 (11), 1574–1581. doi:10.1101/gr.109595.110
- Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. and Chem.* 18 (3), 269–285. doi:10.1016/0097-8485(94)85023-2
- Wootton, J. C., and Scott, F. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. and Chem.* 17 (2), 149–163. doi:10.1016/0097-8485(93)85006-X
- Wright, P. E., and Jane Dyson, H. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16 (1), 18–29. doi:10.1038/nrm3920