bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

MolEvolvR: A web-app for characterizing proteins using molecular evolution and phylogeny

Joseph T Burke[1,2,#], Samuel Z Chen[1,2,#], Lo M Sosinski[1,3,#], John B Johnston[5], Janani Ravi[1*]

[1]Department of Pathobiology and Diagnostic Investigation, Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824; [2]Genomics and Molecular Genetics Undergraduate Program, Michigan State University, East Lansing, MI 48824; [3]Computer Science Undergraduate Program, Michigan State University, East Lansing, MI 48824; [4]Biochemistry and Biotechnology Undergraduate Program, Michigan State University, East Lansing, MI 48824; [5]Center for Genomics-Enabled Plant Sciences, Michigan State University, East Lansing, MI 48824.
[#]Co-primary authors contributed equally, listed alphabetically.
*Corresponding author: janani@msu.edu.

Abstract

Studying proteins through the lens of evolution can help identify conserved features and lineage-specific variants, and potentially, their functions. *MolEvolvR* (http://jravilab.org/molevolvr) is a web-app that enables researchers to run a general-purpose computational workflow for characterizing the molecular evolution and phylogeny of their proteins of interest. The web-app accepts input in multiple formats: protein/domain sequences, homologous proteins, or domain scans. *MolEvolvR* returns detailed homolog data, along with dynamic graphical summaries, *e.g.*, MSA, phylogenetic trees, domain architectures, domain proximity networks, phyletic spreads, and co-occurrence patterns across lineages. Thus, *MolEvolvR* provides a powerful, easy-to-use interface for computational protein characterization.

Keywords

Molecular evolution; Phylogenetics; Domain Architectures; Homology; Protein characterization; Comparative genomics; Web application

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The rate at which molecular or biochemical functions are assigned to proteins greatly lags behind the rate at which new protein families are discovered[1]. This gap impedes identifying and characterizing the complete molecular systems involved in various critical cellular processes such as molecular pathogenesis, antibiotic resistance, or stress response. Several studies[2–16] have demonstrated the power of molecular evolution and phylogenetic analysis in determining molecular functions of such proteins. There are many individual tools[17–36] for protein sequence similarity searches or ortholog detection, delineating co-occurring domains (domain architectures), and building multiple sequence alignments and phylogenetic trees. However, there is a paucity of unified software or web frameworks that effectively integrate these approaches to comprehensively characterize proteins and help discern function by exhaustively identifying all members of relevant protein families (including remote homologs), mapping domains/domain-architectures, and tracing their phyletic spread across lineages.

Here, we present *MolEvolvR*, a web application that provides a streamlined, easy-to-use platform for comprehensively characterizing proteins (**Fig. 1**; accessible at http://jravilab.org/molevolvr). *MolEvolvR* performs homology searches across the tree of life and reconstructs domain architecture by characterizing the input proteins and each of their homologs and presents these results in the context of evolution across the tree of life. The computational evolutionary framework underlying *MolEvolvR* is written using custom R[37–52] and shell scripts. The web application is built with an R/Shiny[39,53,54] framework with additional UI customizations in HTML, Javascript, and CSS, and has been tested on Chrome, Brave, Firefox, and Safari browsers on Mac, Windows, and Linux operating systems.

To illustrate the full range of analysis made possible by *MolEvolvR*, we consider a researcher starting with a protein of unknown function. In Step 1, *MolEvolvR* resolves this query protein into its constituent domains and uses each domain for iterative homology searches[18,19] across the tree of life (>7,000 complete genomes)[55–58]. This divide-and-conquer strategy will identify all proteins similar to each domain, including remote homologs that cannot be found by homology searches using the full-length protein sequences alone. In Step 2, to delineate molecular function, *MolEvolvR* reconstructs the domain architecture by characterizing the protein and each of its homologs by combining: 1) sequence alignment and clustering algorithms for domain detection[17,22,23,59]; 2) profile matching against protein domain and orthology databases[30,31,60–63]; 3) prediction algorithms for signal peptides[31,64], transmembrane regions[31,65–67], cellular localization[31,66], and secondary/tertiary structures[31,68–71]. This analysis results in a detailed molecular characterization of each query protein and its homologs across the three kingdoms. *MolEvolvR* is versatile in many ways. First, the web-app accepts multiple types of queries. Researchers can start their searches using protein/domain sequences of single- or multi-protein operons (*e.g.,* FASTA, NCBI protein accession numbers), homologous proteins (*e.g.,* web or command-line BLAST output), or motif/domain scans (*e.g.,* InterProScan output) [**Fig. 1B**]. Second, it enables tailored analyses to answer a variety of questions (*e.g.*, determining protein features restricted to certain pathogenic groups to discover virulence factors/diagnostic targets) [**Fig. 1C**]. Finally, it provides multiple types of outputs (*e.g.*, complete set of homologs/phylogenetic tree, domain architecture of query protein, most

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

common partner domains) [previews in **Fig. 1A**]. Besides tables and visualizations for each result, *MolEvolvR* also provides graphical summaries that bring these results together in the context of evolution: i) structure-based multiple sequence alignments and phylogenetic trees; ii) domain proximity networks to consolidate results from all co-occurring domains (across homolog domain architectures; iii) phyletic spreads of homologs and their domain architectures; and iv) co-occurrence patterns and relative occurrences of domain architectures across lineages [**Fig. 1A**]. The web-app contains detailed documentation about all these options.

A specific instance of the web-app, applied to study several Psp stress response proteins (present across the tree of life), can be found here: https://jravilab.shinyapps.io/psp-evolution[2]. *MolEvolvR* is a generalized web-server of this web-app. To demonstrate its broad applicability, we have applied the approach underlying *MolEvolvR* to study several systems, including proteins/operons in zoonotic pathogens, *e.g.*, nutrient acquisition systems in *Staphylococcus aureus*[4,5], novel phage defense system in *Vibrio cholerae*[6], surface layer proteins in *Bacillus anthracis*[7], helicase operators in bacteria[8], and internalins in Listeria[9].

Thus, *MolEvolvR* (http://jraviilab.org/molevolvr) is a flexible and powerful interactive web tool that researchers can use to bring molecular evolution and phylogeny to bear on their proteins of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

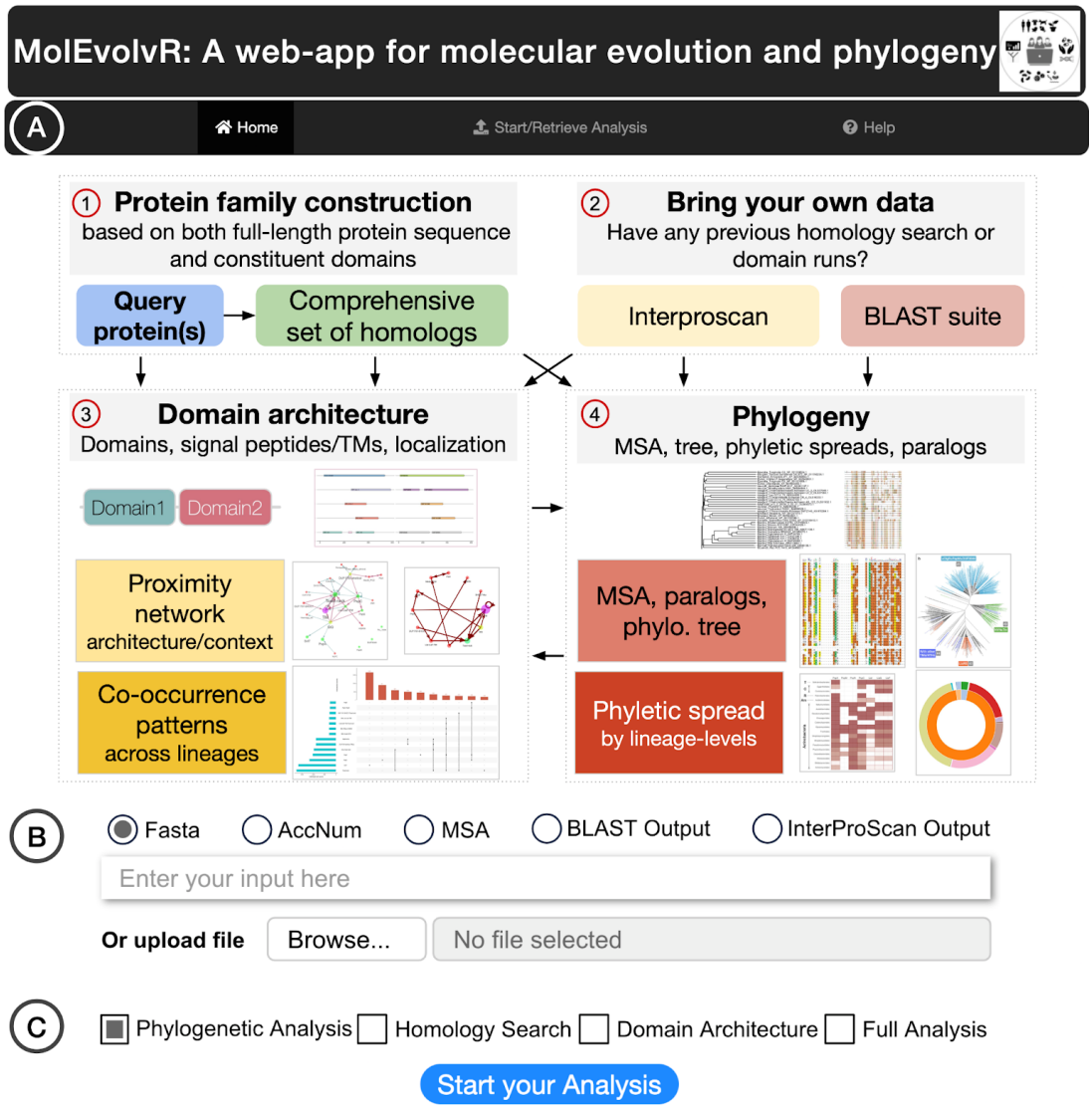


Figure 1. Overview of *MolEvolvR*. A. *MolEvolvR* allows users to start with protein(s) of interest and perform the full analysis (1+3+4), only protein characterization (1+3), only homology searches (1+4), or start with external outputs from BLAST or Interproscan for further analysis, summarization, and visualization (2+3+4). *MolEvolvR* is interactive, queryable, and customizable. **B.** Multiple input options in *MolEvolvR*: FASTA, NCBI accession numbers, pre-computed MSA (in FASTA formats), analysis outputs from an external web or command-line BLAST or InterProScan runs. **C.** The different analysis options available in *MolEvolvR* are based on the chosen input formats in **1A**, **1B**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Material

Supplementary Text

Here, we present a case study with the Lia operon (from *Bacillus subtilis*) involved in lantibiotic stress response, one of the better-studied phage shock protein (PSP) stress response systems[2]. The PSP systems are well-known for having a good mix of conserved components (*e.g.,* PspA) and variant themes (*e.g.,* centering on PspBC or Lia components). It is unclear upfront as to what the underlying building blocks (domains/domain architectures) might be that culminate in the stress response function and what their history is in the broader context of bacterial evolution (phylogeny, phyletic spreads). Therefore, we focus on a detailed characterization of the six Lia proteins, LiaIHGFSR, in this section; the detailed results (analyzed data and graphical summaries) are available here for interactive exploration: jravilab.org/molevolvr/?r=k0PsZM. We used *MolEvolvR* to conduct a comprehensive analysis across the three superkingdoms of life and delineate all occurrences of Lia proteins in terms of their underlying domain architectures and phyletic spreads.

***Homologs*.** We first generated the homologs for each of these six proteins across all completed representative/reference genomes (from bacteria, archaea, and eukaryota) [Data tab, web-app]. The homolog data from these genomes are available in an interactive and queryable format in the *MolEvolvR* web-app with linked NCBI accession numbers, species, lineages, percentage similarities, and several protein- and similarity-search-related parameters [**Fig. S1**]. Best hits by species or lineage can be easily subsetted. The whole data table is available for download in a standard CSV format.

***Domain architectures*.** Next, we determine the domain architectures (based on sequence-structure motifs/domains, disorder predictions, transmembrane regions, signal peptides, and cellular localizations) of the six Lia proteins and each of their homologs. In the 'Domain Architecture' tab, we first summarize our results for individual and across all query proteins [**Fig. S2A**, left]. In addition to the diverse domain architectures, we can also determine their phyletic spreads using a simple popup feature to determine widespread vs. lineage-specific domain architectures [**Fig. S2A**, right]. Using *MolEvolvR*, we can also visualize the diverse domain architectures and cellular localizations (*e.g.,* Pfam, Gene3D, Phobius, MobiDB) of representative homologs of each Lia protein to discover both similar and dissimilar homologs, *e.g.,* ones with novel partner domains, variants with altered localization [**Fig. S2B**]. We summarize our domain findings (by profile database/prediction algorithm) in the 'Network' section of the 'Domain Architecture' tab [**Fig. S2C**]. The proximity network consolidates the findings across proteins (or by protein) by connecting co-occurring domains within proteins by their frequency of occurrence across lineages (nodes, domains; edges, co-occurrence; node size and width of arrows, frequencies). This big picture view, along with the co-occurrence plots [**Fig. S2D**], helps discover new components and their relative frequencies across thousands of homologs. They also lead to serendipitous connections across homologs of the different Lia proteins, *e.g.,* variants of the two-component LiaRS system that carry both the

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

response regulator and histidine kinase domains, as well as predominantly lone domains, *e.g.,* PspA_IM30 [**Fig. S2D**].

Phylogeny. Finally, we used *MolEvolvR* to study the evolution of the Lia proteins [Phylogeny tab, *MolEvolvR*]. We recorded the presence of homologs, by lineage, through interactive sunburst plots [**Fig. S3A**; Phylogeny tab, web-app] and heatmap for all query proteins [**Fig. S3B**; Data tab, web-app]. We generated multiple sequence alignments and phylogenetic trees based on key representatives from diverse species, lineages, and domain architectures [**Fig. S3C–D**, Phylogeny tab, web-app].

Through these comprehensive analyses, we first characterized the proteins encoded by the Lia operon in terms of their domain architectures, with LiaH, a PspA stress response effector protein, two transmembrane proteins LiaI and LiaG, two globular domains of unknown function with Toastrack-like domains in LiaF and LiaG, and a two-component system LiaRS with response regulator/receiver domain and histidine kinase. We discovered several homologs in lineages outside Firmicutes for each of the Lia proteins and domains, including the widespread two-component system, while other domains (PspA, DUF2154, DUF4097) were predominantly present in only Firmicute genomes. We also identified new connections within the more extensive Lia proximity network based on homologs of the LiaRS proteins, with proteins that carried domains from both response regulator and histidine kinase domains (*e.g.,* in proteobacteria, planctomycetes), and rare DUF4097-containing homologs within Firmicutes that carry an additional N-terminal DUF1700 or a second DUF4097 domain.

A more comprehensive analysis of the phage shock protein (PSP) system and its partner domains across the tree of life using a *MolEvolvR*-like approach leading to multiple novel discoveries has been described in a recent article[2]. The PSP system is a great use case to characterize using molecular evolution and phylogeny due to the variety of domain architectures, cellular localizations, and phyletic spreads of each of these protein families and operons. The PSP work, including this Lia operon use case, as well as other recent diverse biological applications[2–11] indicate that the *MolEvolvR* approach is invaluable in characterizing new protein families of interest in the context of evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Figures

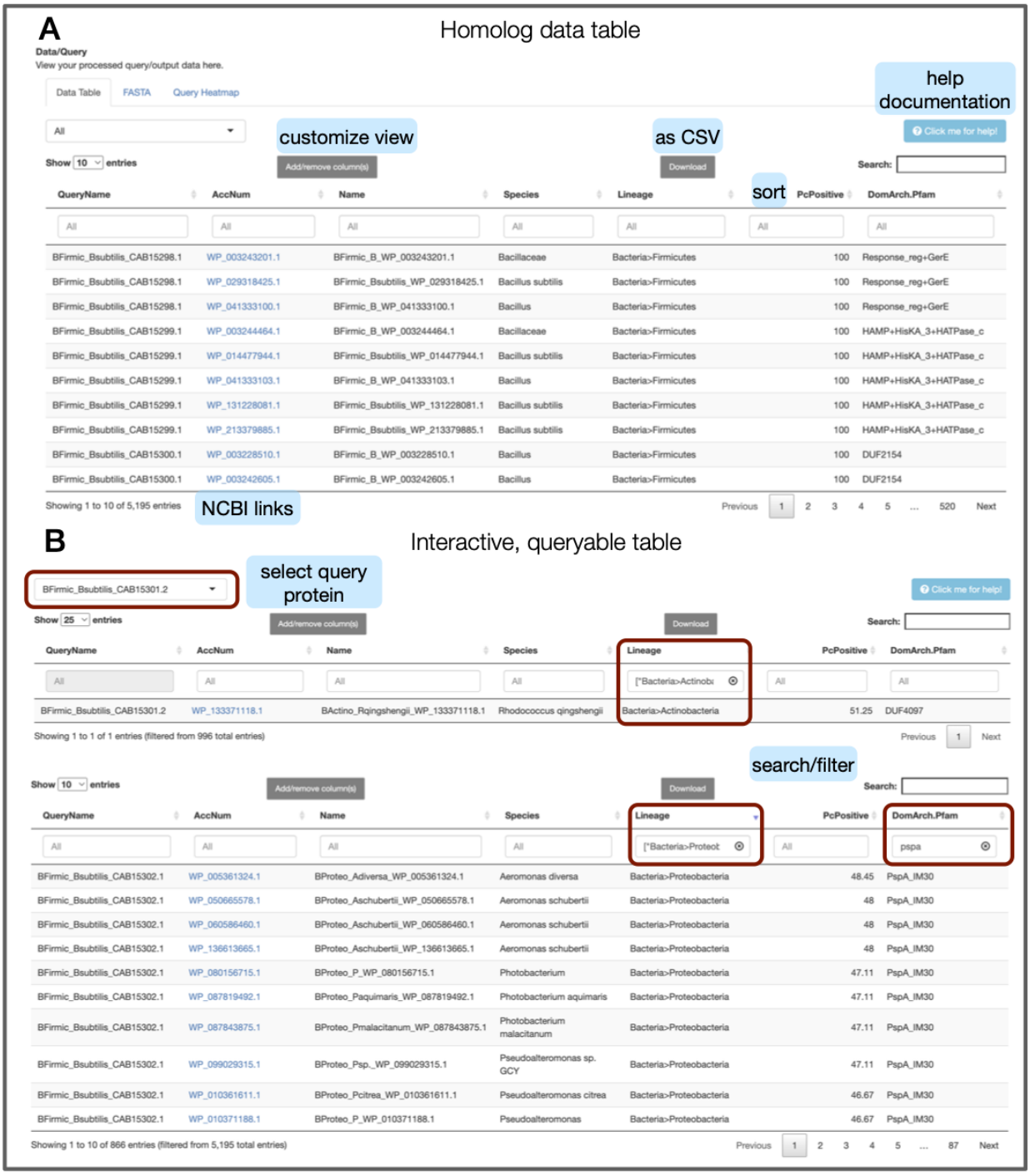


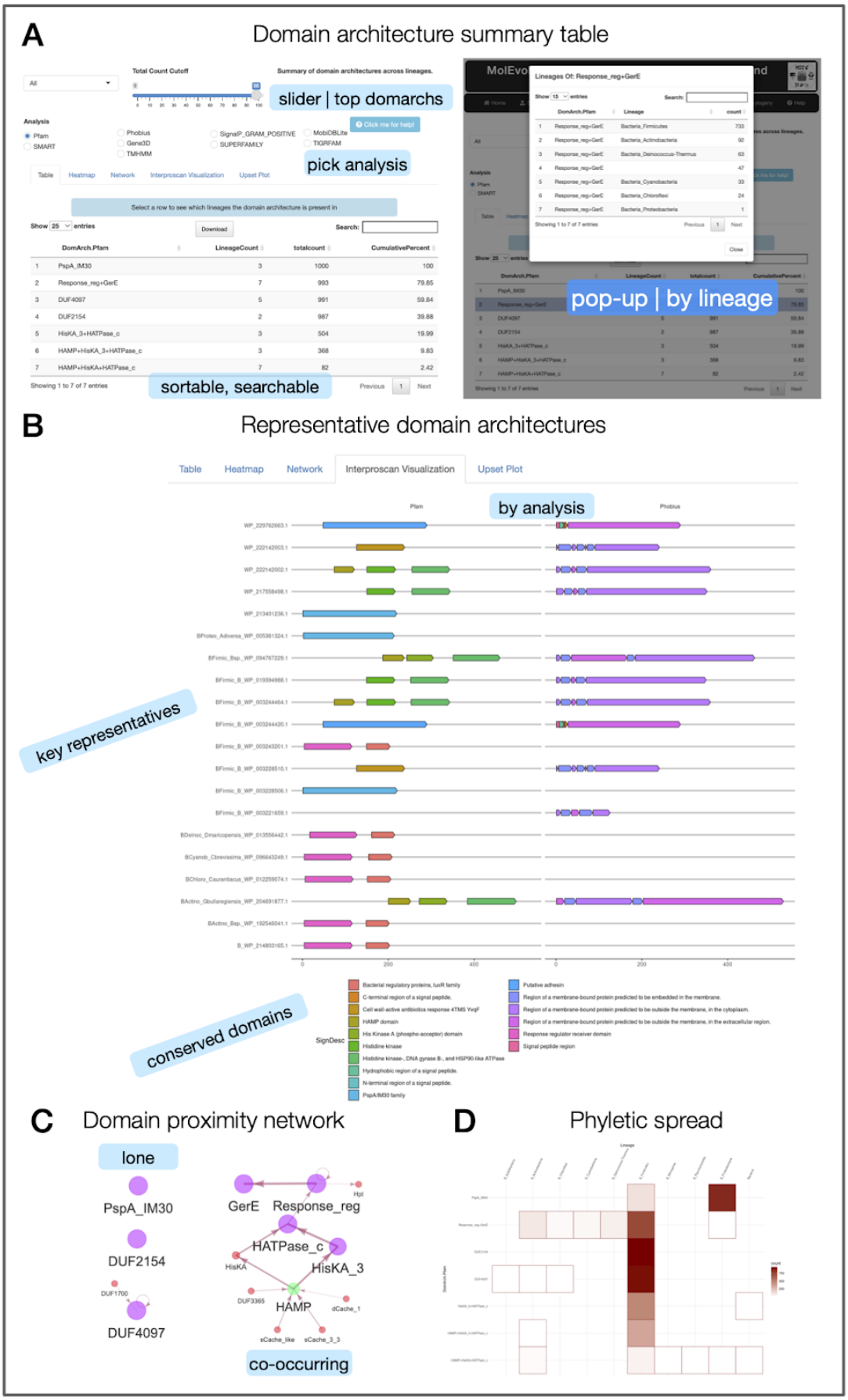
Figure S1. Explore the homologs, their lineages, and domain architectures with MolEvolvR.

Homolog data table with the best hits from all superkingdoms of life (queried across all RefSeq genomes). For each homolog, we tabulate details across all query proteins, including genome, species, lineage, sequence homology (BLAST parameters), and domain architecture information in a sortable, queryable, and interactive manner. The 'Add/remove column(s)'

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

button allows users to access the full list of columns and add/remove column(s) as needed. 'Download' lets the user download the filtered subset of the table or the entire table for further local analysis (as *CSV*). Results across and within individual protein searches can be viewed (using the dropdown menu on the top right). The accession number for each homolog is hyperlinked to its corresponding NCBI protein page. **A** shows a snapshot of the best hits across all query proteins with the default columns. **B** shows examples of the **Interactive, queryable table** with homologs being filtered based on specific lineages (*e.g.,* actinobacteria) AND *top*: query protein (BFirmic_Bsubtilis_CAB15301.2) or *bottom*: domains (*e.g.,* PspA). *Light blue boxes with text are not part of the screenshot; they have been added to highlight key features.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Figure S2. Domain architectures of query proteins and their homologs using *MolEvolvR*.

A. Domain architecture summary table with phyletic spreads. The table shows the top (most predominant) domain architecture by query protein (or across all queries, as in this case) with the frequency of occurrence and lineages in which they occur. The slider allows users to pick the top hits, representing 98% of all homologs in this case. The second snapshot shows the popup that appears when each domain architecture row is clicked; it elaborates on the 'LineageCount' by showing the frequencies of occurrence by individual lineage for the selected domain architecture. **B. Representative domain architectures.** The figure shows cartoon depictions of representative domain architectures (Pfam) and cellular localizations (Phobius) of the 6 query proteins. The Pfam and Phobius annotations for each domain prediction (by colour) are shown in the legend. **C. Domain proximity network.** The network captures co-occurring domains within the top 98% of the homologs of all the 'query' Psp members and their key partner domains (after sorting by decreasing frequency of occurrence). The size of the nodes (domains) and width of edges (co-occurrence of domains within a protein) are proportional to the frequency of their occurrence across homologs. The query domains (original proteins/domains of interest) and other commonly co-occurring domains are indicated in red and grey. The complete network, as well as the domain-centric ones, are available on the web-app. **D. Phyletic spread** of the predominant domain architectures across query proteins. The heatmap shows the presence/absence of homologs of all query proteins across key lineages (columns) for each predominant domain architecture (rows). The colour gradient indicates the highest number of homologs in a particular lineage. The heatmap gives the full picture. *Rows*: Top domain architectures across all homologs. *Columns*: The major archaeal, bacterial, eukaryotic, and viral lineages with representative sequenced genomes. *Blue boxes with text are not part of the screenshot; they have been added to highlight key features.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

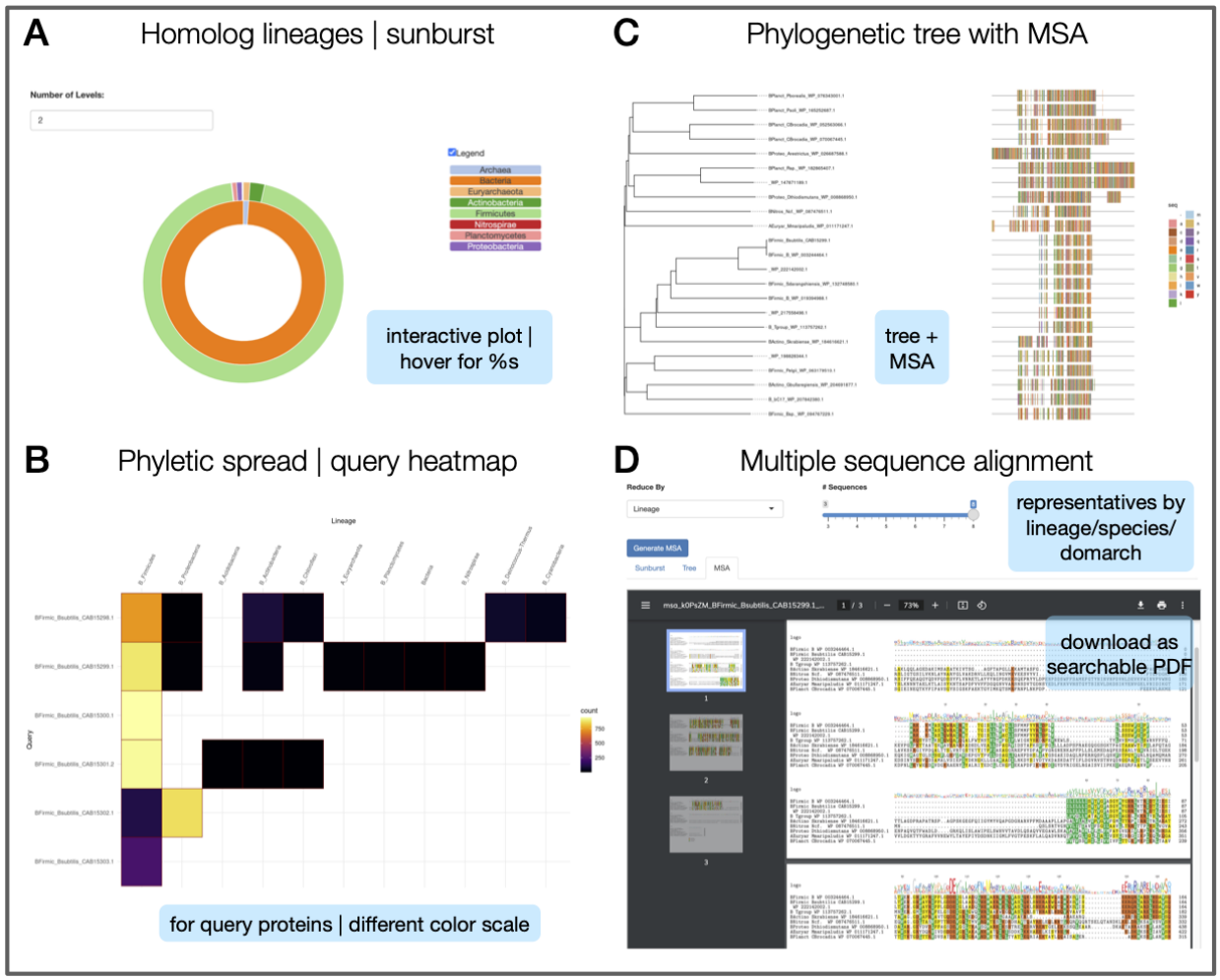


Figure S3. Phyletic spreads and phylogeny of query proteins using *MolEvolvR*.

A. Lineages of homologs. Sunburst plot shows the phyletic spread of all the homologs. The legend shows the lineages (inner ring, kingdoms; outer ring, phyla) that carry homologs. Note: The sunburst plots only display lineages of >0.1% fraction of total proteins. **B. Phyletic spread** of the homologs by query protein. The heatmap shows the presence/absence of homologs across key lineages (columns) for each query (rows). The colour gradient indicates the highest number of homologs in a particular lineage. The heatmap gives the whole picture. *Rows*: Query proteins are queried against all sequenced and completed genomes across the three major kingdoms of life. *Columns*: The major archaeal, bacterial, eukaryotic, and viral lineages with representative sequenced genomes. **C.** The **multiple sequence alignment** is overlaid on the **phylogenetic tree**. In the tree generated for LiaS (BFirmic_Bsubtilis_CAB15299.1), each leaf (protein) is named with its kingdom (*e.g.,* 'B' for bacteria), phylum (first six letters, *e.g.,* 'actino' for actinobacteria), Genus, species (represented as 'Gspecies,' *e.g.,* 'Bsubtilis' for *Bacillus subtilis*), and NCBI protein accession number (*e.g.,* CAB15298.1), resulting in a uniquely identifiable name for the protein, 'KPhylum_Gspecies_AccNum,' *e.g.,* BFirmic_Bsubtilis_CAB15299.1. *Key*: the colours in the multiple sequence alignment depiction correspond to different amino acids. **D.** Snapshot of the **multiple sequence alignment** of

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

representative homologs of LiaS (BFirmic_Bsubtilis_CAB15299.1) by lineage. The generated MSA, with representative homologs from each lineage, species, or domain architectures, is available to users as a downloadable searchable PDF. *Light blue boxes with text are not part of the screenshot; they have been added to highlight key features.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Acknowledgments

We would like to thank members of the JRaviLab (Elliot Majlessi, Ethan Wolfe, Karn Jongnarangsin, Kewalin Samart) for testing the web-app at various stages and providing the authors with several iterations of constructive feedback. We are also extremely grateful to Krishnan Raghunathan, Premal Shah, Arjun Krishnan, and members of the Krishnan lab (Kayla Johnson, Nathaniel Hawkins) for early feedback on *MolEvolvR* and the manuscript. We have benefited from several diverse use cases and challenges brought to us by our collaborators that helped fine-tune the functionality of the web-app, thanks to collaborations with L Aravind, Neal Hammer, Christopher Waters, Antonella Fioravanti, Jonathan Hardy, Kayla Conner, Christina Stallings, Helen Blaine, and Stephanie Shames. We appreciate timely help from Kellen Reason, our current system administrator, for maintaining the backend of our web-server.

Funding

We would like to thank our funding sources: Endowed Research Funds from the College of Veterinary Medicine, Michigan State University, NSF-funded BEACON funding support, and Michigan State University start-up funds awarded to JR; NSF-funded REU-ACRES summer scholarship to SZC.

Author Contributions

JR conceived the study; JR designed the study; JTB, SZC, LMS, and JR acquired the data, performed all the analyses, and made the figures and tables, specifically JTB, SZC, LMS, JR wrote the backend code, JTB, SZC, JR built the R/Shiny web-app, JBJ set up the server backend for the web-app; LMS and JR wrote the first draft of the manuscript; JTB, LMS, and JR revised the manuscript.

Data Availability and Reuse

All the data, analyses, and visualizations are available in our interactive and queryable web application: http://jravilab.org/molevolvr. Text, figures, and the webapp are licensed under Creative Commons Attribution CC BY 4.0.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

References

1. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
2. Ravi, J. *et al.* The Phage-shock-protein (PSP) Envelope Stress Response: Discovery of Novel Partners and Evolutionary History. *bioRxiv [Preprint]* 2020.09.24.301986 (2020) doi:10.1101/2020.09.24.301986.
3. Ravi, J., Anantharaman, V., Aravind, L. & Gennaro, M. L. Variations on a theme: evolution of the phage-shock-protein system in Actinobacteria. *Antonie Van Leeuwenhoek* **111**, 753–760 (2018).
4. Lensmire, J. M. *et al.* The Staphylococcus aureus Cystine Transporters TcyABC and TcyP Facilitate Nutrient Sulfur Acquisition during Infection. *Infect. Immun.* **88**, (2020).
5. Lensmire, J. M. *et al.* The glutathione import system satisfies the Staphylococcus aureus nutrient sulfur requirement and promotes interspecies competition. *bioRxiv [Preprint]* 2021.10.26.465763 (2021) doi:10.1101/2021.10.26.465763.
6. Severin, G. B. *et al.* A Broadly Conserved Deoxycytidine Deaminase Protects Bacteria from Phage Infection. *bioRxiv [Preprint]* 2021.03.31.437871 (2021) doi:10.1101/2021.03.31.437871.
7. Ravi, J. & Fioravanti, A. S-layers: The Proteinaceous Multifunctional Armors of Gram-Positive Pathogens. *Front Microbiol* **12**, 663468 (2021).
8. Blaine, H. C., Burke, J. T., Ravi, J. & Stallings, C. L. DciA helicase operators exhibit diversity across bacterial phyla. *bioRxiv [Preprint]* 2022.01.24.477630 (2022) doi:10.1101/2022.01.24.477630.
9. Conner, K. N., Burke, J. T., Ravi, J. & Hardy, J. W. Novel Internalin P homologs in Listeria.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

bioRxiv [Preprint] 2022.01.19.476994 (2022) doi:10.1101/2022.01.19.476994.

10. Manganelli, R. & Gennaro, M. L. Protecting from Envelope Stress: Variations on the Phage-Shock-Protein Theme. *Trends Microbiol.* **25**, 205–216 (2017).

11. Datta, P. *et al.* The Psp system of Mycobacterium tuberculosis integrates envelope stress-sensing and envelope-preserving functions. *Mol. Microbiol.* **97**, 408–422 (2015).

12. Kaur, G., Burroughs, A. M., Iyer, L. M. & Aravind, L. Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity. *Elife* **9**, (2020).

13. Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V. & Koonin, E. V. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5307–E5316 (2018).

14. Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol. Direct* **8**, 15 (2013).

15. Koonin, E. V. Systemic determinants of gene evolution and function. *Mol. Syst. Biol.* **1**, 2005.0021 (2005).

16. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).

17. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

18. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).

19. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

20. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

21. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).

22. Magis, C. *et al.* T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol. Biol.* **1079**, 117–129 (2014).

23. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

24. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz795.

25. Gumerov, V. M. & Zhulin, I. B. TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa243.

26. Klimchuk, O. I. *et al.* COGNAT: a web server for comparative analysis of genomic neighborhoods. *Biol. Direct* **12**, 26 (2017).

27. Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat Protoc* **14**, 3013–3031 (2019).

28. Adebali, O. & Zhulin, I. B. Aquerium: A web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree. *Proteins* **85**, 72–77 (2017).

29. Persson, E., Kaduk, M., Forslund, S. K. & Sonnhammer, E. L. L. Domainoid: domain-oriented orthology inference. *BMC Bioinformatics* **20**, 523 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

30. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).

31. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

32. Altenhoff, A. M. *et al.* OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**, D373–D379 (2021).

33. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics* **35**, 149–151 (2019).

34. Forslund, K., Henricson, A., Hollich, V. & Sonnhammer, E. L. L. Domain tree-based analysis of protein architecture evolution. *Mol. Biol. Evol.* **25**, 254–264 (2008).

35. Forslund, S. K., Kaduk, M. & Sonnhammer, E. L. L. Evolution of Protein Domain Architectures. *Methods Mol. Biol.* **1910**, 469–504 (2019).

36. Haider, C., Kavic, M. & Sonnhammer, E. L. L. TreeDom: a graphical web tool for analysing domain architecture evolution. *Bioinformatics* **32**, 2384–2385 (2016).

37. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).

38. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, PBC., 2020).

39. Allaire, J. J. *et al. rmarkdown: Dynamic Documents for R*. (2020).

40. JJ Allaire, T., Yihui Xie, R. Foundation, Hadley Wickham, Journal of Statistical Software, RStudio, Ramnath Vaidyanathan, Association for Computing Machinery, Carl Boettiger, Elsevier, Karl Broman, Kirill Mueller, Bastiaan Quast, Randall Pruim, Ben Marwick, Charlotte Wickham, Oliver Keyes, Miao Yu, Daniel Emaasit, Thierry Onkelinx, Alessandro Gasparini, Marc-Andre Desautels, Dominik Leutnant, MDPI & Francis, C. D., Oguzhan Ögreden, Dalton

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Hance, Daniel Nüst, Petter Uvesten, Elio Campitelli, John Muschelli, Alex Hayes, Zhian N. Kamvar, Noam Ross, Robrecht Cannoodt, Duncan Luguern, David M. Kaplan, Sebastian Kreutzer, Shixiang Wang, Jay Hesselberth, Alfredo Hernández. *Article Formats for R Markdown*. (2020).

41. Xie, Y., Allaire, J. J. & Grolemund, G. *R Markdown: The Definitive Guide*. (Chapman and Hall/CRC, 2018).

42. Müller, K. *here: A Simpler Way to Find Your Files*. (2017).

43. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

44. Jake Conway, N. G. *A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. (2019).

45. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983–1992 (2014).

46. Perry, M. *Flexible Heatmaps for Functional Genomics and Sequence Features*. (2016).

47. Fellows, I. *Word Clouds*. (2018).

48. Dawei Lang, G. C. *Create Word Cloud by 'htmlwidget*. (2018).

49. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

50. Enrico Bonatesta, U. B., Christoph Horejs-Kainrath. *Multiple Sequence Alignment*. (2020).

51. Hao ZhuThomas Travison, D. M., Timothy Tsai, Will Beasley, Yihui Xie, GuangChuang Yu, Stéphane Laurent, Rob Shepherd, Yoni Sidi, Brian Salzer, George Gui, Yeliang Fan. *Construct Complex Table with 'kable' and Pipe Syntax*. (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

52. Mike Bostock, C. Y., Kerry Rodden, Kevin Warne, Kent Russell, Florian Breitwieser. *Sunburst 'Htmlwidget'*. (2020).

53. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. *shiny: Web Application Framework for R*. (2019).

54. Winston Chang, R. C. T., Joe Cheng, JJ Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation, jQuery contributors, jQuery UI contributors, Mark Otto, Jacob Thornton, Bootstrap contributors, Twitter, Inc, Alexander Farkas, Scott Jehl, Stefan Petre, Andrew Rowls, Dave Gandy, Brian Reavis, Kristopher Michael Kowal, es5-shim contributors, Denis Ineshin, Sami Samhuri, SpryMedia Limited, John Fraser, John Gruber, Ivan Sagalaev. *Web Application Framework for R*. (2020).

55. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36-42 (2013).

56. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

57. Tatusova, T. *et al.* Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* **43**, D599-605 (2015).

58. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).

59. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).

60. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).

61. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).

62. Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Microbial genome analysis: the COG approach. *Brief. Bioinformatics* **20**, 1063–1070 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.461833; this version posted February 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

63. Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. CDART: protein homology by domain architecture. *Genome Res.* **12**, 1619–1623 (2002).

64. Nielsen, H. Predicting Secretory Proteins with SignalP. *Methods Mol. Biol.* **1611**, 59–73 (2017).

65. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

66. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35**, W429-432 (2007).

67. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

68. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244-248 (2005).

69. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197-201 (2008).

70. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).

71. C, Y., J, L., P, C., I, S. & C, O. The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic acids research* vol. 39 https://pubmed.ncbi.nlm.nih.gov/21646335/ (2011).