OrthoGather: a local platform for orthology-based proteome and proteomics comparisons and Gene Ontology enrichment

Carlos Vivas-Rodríguez[1,2,3], David Gomez Matallanas[3,4], Colm J. Ryan[2,3], Siobhán McClean[1,2], Olivier Dennler[2,3*], Joanna Drabinska[1,2*]

1. School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland
2. UCD Conway Institute of Biomedical and Biomolecular Research, Dublin, Ireland
3. School of Medicine, University College Dublin, Dublin, Ireland
4. Systems Biology Ireland, University College Dublin, Dublin, Ireland

*** Corresponding authors:** Joanna Drabinska (joanna.drabinska.pro@gmail.com); Olivier Dennler (olivier.dennler@gmail.com)

Abstract

Comparative proteomic analysis can reveal pathways that are regulated by the same stimulus in different species, such as common responses to antibiotics across different bacteria. These data often derive from differential protein expression analyses or from curated protein sets associated with specific biological processes. Functional annotations are key to such comparisons but are of variable quality for different species. A significant number of proteins are poorly or inconsistently annotated, particularly in prokaryotes and non-model eukaryotes, which limits biological interpretation and complicates comparative studies. Although orthology inference provides a robust framework to address this issue, current tools require technical expertise, command-line execution and manual handling of complex outputs, which creates barriers for researchers without any computational training.

We developed *OrthoGather*, a locally hosted web application that streamlines comparative proteomic analysis by integrating groups of homologous proteins across species (orthogroup inference), interactive visualisation, and Gene Ontology (GO) enrichment into a unified platform. *OrthoGather* generates comprehensive, user-friendly, and easy-to-interpret outputs and supports GO enrichment by leveraging functional annotations available in any member of an orthogroup**,** enabling functional inference even when individual species are poorly annotated. Its flexible design enables cross-species exploration of conserved and unique orthogroups, while user-defined protein sets are evaluated against appropriate background reference sets, leveraging orthogroup relationships to reveal functional patterns, even in poorly annotated proteomes.

Introduction

The incomplete functional annotation of proteins presents a major challenge in proteomics, particularly when researchers seek to interpret the biological significance of protein sets derived from experimental studies. This issue is especially challenging when working with prokaryotes or non-model eukaryotes, where incomplete or inconsistent annotations are commonplace (Linard et al. 2021; Warren et al. 2010; Chaudhari et al. 2024; Lobb et al. 2020; Cantalapiedra et al. 2021b). This variability in annotations also complicates cross-species comparisons of whole proteomes and experimental proteomic datasets, limiting the identification of functionally equivalent proteins, and the identification of shared or unique responses to stimuli (UniProt Consortium 2025; Huerta-Cepas et al. 2019; Wattam et al. 2017).

Orthology inference offers a robust strategy to address this limitation by identifying homologous, functionally related proteins across species. Tools such as OrthoFinder enable the identification of conserved and species-specific gene sets through accurate orthology assignment, clustering evolutionarily related proteins into orthogroups (Emms and Kelly 2015, 2019). However, the use of such tools is often limited by technical barriers, including the need for command-line execution, manual data preparation, and complex, hard-to-interpret outputs. These technical demands limit the accessibility for researchers without computational expertise.

To address these issues, we developed *OrthoGather*, a locally hosted web application that streamlines comparative proteomic analysis. It retrieves reference proteomes from the UniProt database (UniProt Consortium, 2025), executes OrthoFinder locally, and generates browser-based, downloadable figures and summary statistics. Functional interpretation is further supported by Gene Ontology (GO) enrichment, enabling the analysis of user-defined protein sets in a comparative, cross-species context. Sharing GO annotations across orthogroups facilitates the identification of differentially expressed pathways, even in poorly annotated species and strains. An example of use of *OrthoGather* would be the analysis of protein sets derived from differential proteomics experiments, where changes in protein abundance are compared across conditions in distinct species. While orthogroup inference relies on reference proteomes, these experiments generate subsets of proteins associated with specific responses, whose

comparative analysis across species enables the identification of conserved and species-specific functional patterns.

Workflow and Output Overview

OrthoGather integrates comparative proteome analysis with GO enrichment, streamlining the workflow from data acquisition to interpretation (Fig 1). An example application of this tool and illustration of the workflow, based on a differential proteomics dataset derived from a study examining the response of *Mycobacterium smegmatis* (*Msm*) to sub-lethal concentrations of rifampicin (Giddey et al. 2017), is provided in the Supplementary Materials S3, and used throughout this manuscript.

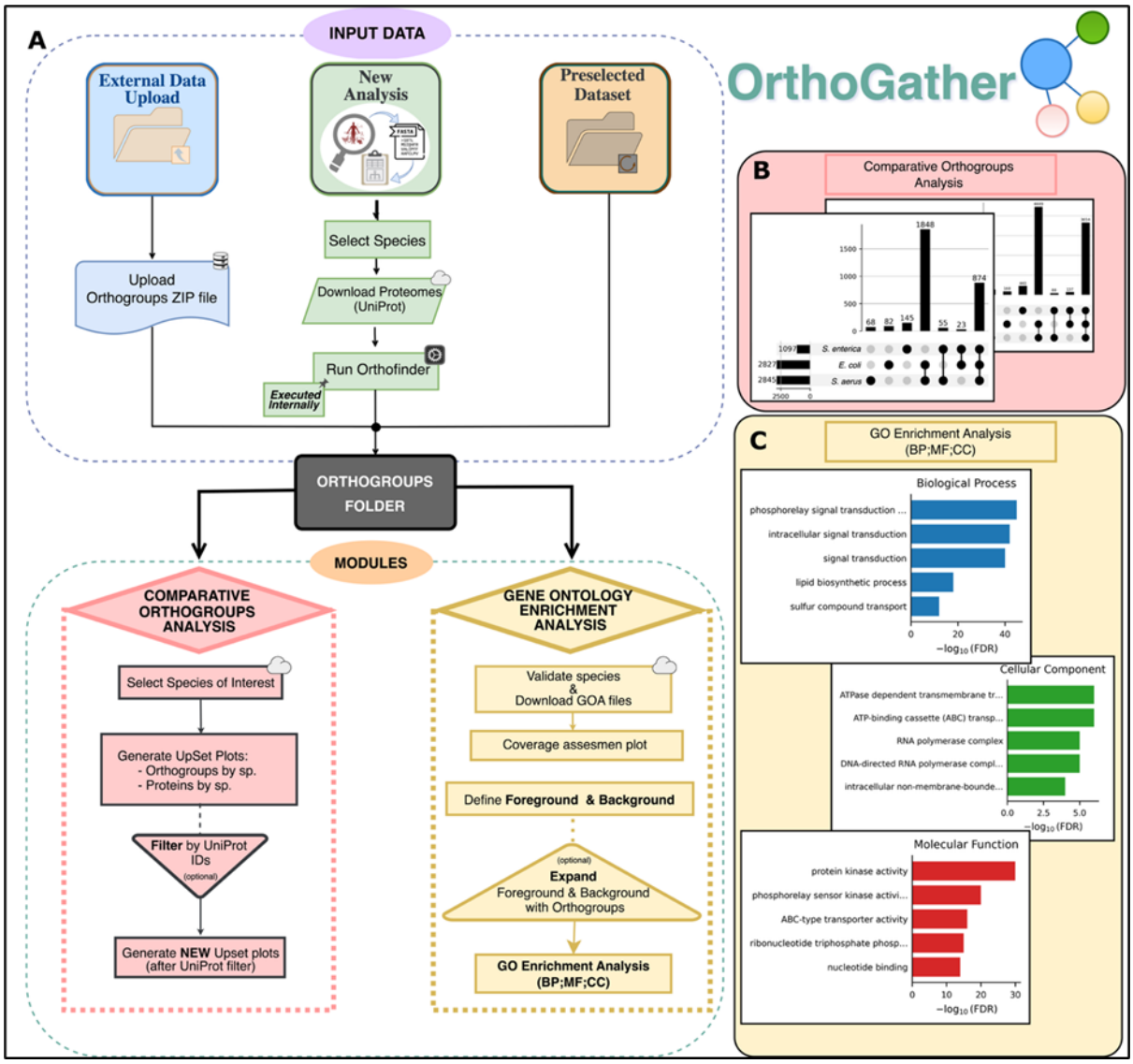

Figure 1. Overview of the *OrthoGather* workflow and interface. (A) Schematic representation of the *OrthoGather* workflow. All three input modes: *External Data Upload*, *New Analysis*, and *Preselected Dataset* are converged in the *Orthogroups* folder. The resulting data can be explored through two independent modules: *Comparative Orthogroup Analysis* and *Gene Ontology Enrichment*. **(B)** Comparative Orthogroup Analysis module illustrating representative UpSet plots that summarise orthogroup overlap and protein distribution across the selected species. **(C)** Representative bar plots showing enriched GO terms across Biological Process (top), Molecular Function (middle), and Cellular Component (bottom) generated through the *Gene Ontology enrichment analysis*.

Three input modes are supported by *OrthoGather* (Fig 1A):

(i) *New Analysis*: A predictive search interface (species name or UniProt proteome ID) allows the selection of proteomes available in the UniProt database. The chosen reference proteomes are automatically downloaded in FASTA format, and OrthoFinder is executed to infer the orthogroups that form the basis for the subsequent analyses.

(ii) *Preselected Dataset*: A fixed example dataset containing 47 proteomes from 17 bacterial species is distributed with the *OrthoGather* installation (Supplementary Materials S2). These species include clinically relevant organisms studied in the context of cystic fibrosis-associated chronic infections and/or antimicrobial resistance. The dataset enables users to quickly familiarise themselves with the platform without waiting for downloads or running external computations.

(iii) *External Data Upload***:** Any OrthoFinder results can be uploaded directly as a compressed *.zip* archive containing the *Orthogroups* directory, enabling the reuse of previously completed analyses.

Regardless of the input mode, *OrthoGather* centralises processing in the *Orthogroups* folder and automatically removes temporary and auxiliary files from the OrthoFinder output to optimise disk usage. Subsequently, users can perform:

1. Comparative Orthogroup Analysis

This module allows users to explore orthogroup composition across species using two successive filtering phases.

(i) *Species filtering phase*. A subset of two or more species is selected from the dataset. *OrthoGather* then retains only those orthogroups containing at least one protein from the selected species, excluding orthogroups formed exclusively by non-selected taxa. This step defines the later analysis framework for the chosen species set. Two

124 UpSet plots are generated (Fig 1B): the first shows the distribution of orthogroups across
125 species combinations, highlighting shared and unique groups; the second displays the
126 number of proteins represented in each intersection, helping to interpret overlaps and
127 relative species contributions. To illustrate the workflow, we used the *New Analysis* input
128 mode with the proteomes of *Msm*, *Acinetobacter baumannii, Burkholderia cenocepacia,*
129 *Escherichia coli, Mycobacterium abscessus and Pseudomonas aeruginosa* to assess
130 whether protein families affected by rifampicin exposure are conserved or species-
131 specific (Supplementary File S1, Fig. S1-S2; Supplementary File S4)

132 **(ii)** *Protein filtering phase*. The analysis can be further narrowed by providing a
133 custom list of UniProt identifiers (UniProt IDs). *OrthoGather* identifies orthogroups that
134 include at least one of the provided protein IDs and produces a new set of UpSet plots
135 (Fig 1B), reflecting orthogroups distribution within this filtered subset. To showcase this,
136 we used the list of 598 proteins differentially expressed following *Msm* rifampicin
137 treatment (Supplementary file S3), showing the orthogroup distribution across the six
138 species, identifying shared and species-specific orthogroups at both orthogroup and
139 protein levels (Supplementary File S1, Fig S3–S4; Supplementary File S4).

140 All results, including the data used to generate the plots, can be exported in Excel format.
141 Each sheet corresponds to one species combination represented in the UpSet plots and
142 contains the list of orthogroups together with the UniProt identifiers of proteins belonging
143 to the specific orthogroup in each species, enabling identification of homologous
144 proteins across selected species, even when their names differ. Example files and
145 representative figures are provided in the Supplementary Materials (Supplementary File
146 S1, Fig S1–S4; Supplementary File S4).

147 **2. Gene Ontology Enrichment Analysis**

148 This module enables functional interpretation of the generated orthogroups (both filtered
149 and unfiltered), provided that the included species have annotations in the Gene
150 Ontology Annotation (GOA) repository (Huntley et al. 2015). For example, GO enrichment
151 can be applied to identify GO terms overrepresented among proteins responding to
152 rifampicin treatment in *Msm*. The workflow consists of three main phases:

153 **(i)** *Species validation phase*. *OrthoGather* automatically checks which of the
154 analysed species have corresponding GOA files by comparing species names against a

155 local reference table of taxon identifiers. Only species with available GO annotations are
156 retained for subsequent steps, ensuring that the analysis is based on annotated proteins
157 and avoiding statistical biases introduced by species without annotation. Manual entry
158 of a taxon ID is possible if required. In the analysed example, all the selected species
159 were automatically matched to their corresponding GOA files without requiring manual
160 taxon ID entry.

161 **(ii)** *Functional coverage phase*. For each orthogroup, *OrthoGather* quantifies the
162 proportion of proteins that are associated with at least one GO term relative to the total
163 number of proteins in the group. This provides a measure of the completeness of GO
164 annotation across the dataset, generating quantitative information on GO annotation
165 coverage across orthogroups, which is visualised as summary distributions and recorded
166 in downloadable output files. The distribution of coverage values is displayed as a
167 histogram and a boxplot, allowing the overall annotation level to be assessed before
168 performing the enrichment. Low coverage suggests that additional well-annotated
169 species could be included to improve downstream results and including functional
170 interpretation. In the analysed example, *E. coli* was included in the analysis due to its
171 high-quality GO annotation, serving as a well-annotated reference and resulting in a
172 median GO annotation coverage of 70.52% across orthogroups with at least one
173 annotated protein (Supplementary File S1, Fig S5; Supplementary File S4).

174 **(iii)** *Enrichment phase*. Two protein sets must be defined: *foreground*, containing the
175 proteins of interest to be evaluated for enrichment, and *background*, against which the
176 foreground is evaluated. The background can be defined either as a custom UniProt ID
177 list or as all proteins from the species with available GOA files. In both cases, the *Include*
178 *Complete Orthogroups* option expands each set to include all proteins from orthogroups
179 containing at least one selected entry, using annotated orthologs to infer function for
180 unannotated proteins within the same orthogroup. For example, the *foreground* was
181 defined as a set of 598 proteins differentially expressed in *Msm* following rifampicin
182 exposure. The *background* typically represents the complete set of proteins detected in
183 the experiment, e.g. 3,246 proteins detected through mass spectrometry in the case of
184 *Msm* (Supplementary File S3). The analysis is conducted using the GOATOOLS library
185 (Klopfenstein et al. 2018), applying Fisher's exact test to determine whether a GO term
186 occurs more frequently in the foreground than expected from the background. Multiple

testing is performed using the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg,1995). Significant results are displayed as bar plots (Fig 1C) separated by GO namespace, with adjustable significance thresholds, term depth, and maximum number of displayed terms. Detailed results are also provided in an Excel table, listing for each enriched term the identifier, description, background and foreground counts, raw p-value, and FDR-adjusted p-value. In the analysed example, GO enrichment analysis was performed using a threshold of FDR<0.001 and two different GO hierarchy depths. Both analyses identified significantly overrepresented terms across the Biological Process, Molecular Function and Cellular Component namespaces, highlighting enrichment of functions related to signal transduction, metabolic processes and transport activities, as well as cellular components associated with membrane structures and protein complexes (Supplementary File S1, Fig S6–S7; Supplementary Files S4). Importantly, *OrthoGather* allows full customization of the analysis parameters, including p-value threshold**,** ontology depth, and maximum number of displayed terms, enabling users to tailor the enrichment output according to the level of detail and required statistical stringency.

Implementation

OrthoGather is developed in Python 3.7.12 and uses Flask 2.2.5 as the backend framework (Grinberg 2018)**,** with HTML and JavaScript for the frontend. Interaction is supported through all modern web browsers without the need for additional client-side installation. The application follows a modular architecture composed of four main components corresponding to the main stages of the workflow (Fig. 1):

1. Proteome selection: predictive search for reference and non-reference proteomes from UniProt (UniProt Consortium 2025), using a locally stored JSON index, followed by automatic download in FASTA format from the UniProt FTP service.

2. Orthogroup inference: OrthoFinder v3.0.1 (Emms and Kelly 2019, 2015; Emms et al. 2025) is executed internally via Python's subprocess module using default parameters, retaining only the Orthogroups directory to optimize storage usage.

3. Orthogroup filtering and exploration: orthogroups can be filtered by selected species and, optionally, by a user-defined list of UniProt identifiers. This module supports

visualization using matplotlib 3.2.2 (Hunter 2007) and UpSet plots (Lex et al. 2014), with export of results in CSV and Excel formats.

4. **Functional analysis using Gene Ontology**: GOA files are obtained and processed from the European Bioinformatics Institute (EBI) (Huntley et al. 2015). This module enables definition of foreground and background sets with optional expansion to include all members of relevant orthogroups, and performs enrichment analysis using GOATOOLS 1.4.12 (Klopfenstein et al. 2018).

OrthoGather is designed to run locally on macOS and Linux systems compatible with OrthoFinder and can be executed on Windows via virtualized environments or the Windows Subsystem for Linux. The platform is openly available on GitHub with installation instructions (Supplementary File S5), documentation, and a requirements.txt file listing all dependencies. An internet connection is required only for retrieving proteomes and GO annotations.

Conclusion

This work introduces a novel tool that addresses a longstanding gap in the bioinformatics toolkit by providing an integrated and accessible platform for comparative and functional proteome analysis. *OrthoGather* empowers researchers by offering a robust and user-friendly framework that enables the complete workflow, from proteome selection to functional enrichment, to be executed locally and reproducibly without requiring advanced programming skills. Beyond the sample dataset presented, *OrthoGather* can be applied to a wide range of biological questions. This may include the analysis of differentially expressed proteins from proteomics experiments, the comparison of proteomic responses across species, and the exploration of conserved and species-specific functions in both well-annotated and poorly annotated organisms. In addition to proteomic analysis, it can be applied to transcriptomic datasets, aid whole genome sequencing analysis, and allow identification of orthologs of specific proteins or genes across species. By leveraging orthology relationships, the platform enables functional inference in under-annotated species through the integration of information from well-annotated reference proteomes, facilitating hypothesis generation and guiding downstream experimental validation.

Acknowledgements

This work was supported by the University College Dublin Conway Institute Director's Strategic Awards For Thematic Research 2023/2024. Colm Ryan and Olivier Dennler were funded by Research Ireland [20/FFP-P/8641] and Siobhán McClean and Joanna Drabinska were funded by Research Ireland [20/FFP-P/8717]. The authors would like to thank Dr Niamh Duggan and Ciarán Carey for testing *OrthoGather* with differential proteomics datasets during its development. We would also like to thank members of the UCD Cancer Data Lab, Dr Metin Yazar and Narod Kebabci for their technical assistance.

Availability and implementation:

Source code, documentation, and tutorials are available at https://github.com/CarlosVivasR/OrthoGather.

Supplementary information:

S1. Supplementary Material information. Detailed specification of all supplementary files provided including example dataset analysis outputs.

S2. Table_S1. List of reference proteomes available within *OrthoGather* in the *Preselected dataset*, including taxonomic information and database identifiers. This dataset is distributed with the software and is used throughout the manuscript to illustrate the *OrthoGather* workflow.

S3. Data_S1. Foreground and background protein sets used for the Gene Ontology enrichment example analysis described in the main text.

S4. Data_S2.zip Output files generated for the example analysis using six representative bacterial species selected from the preselected dataset. This dataset includes orthogroup assignment files produced by OrthoFinder (Orthogroups.tsv, Orthogroups.txt, Orthogroups.GeneCount.tsv and Orthogroups_UnassignedGenes.tsv), data underlying the Comparative Orthogroup Analysis (UpSet plots), and Gene Ontology annotation coverage and enrichment results reported as Excel tables.

S5. Data_S3. Installation guide and configuration instructions for local deployment of the *OrthoGather* platform, including software dependencies and execution requirements.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society: Series B (Methodological) 57 (1): 289–300.

Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021b. "EggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38 (12): 5825–29.

Chaudhari, Jyoti Kant, Shubham Pant, Richa Jha, Rajesh Kumar Pathak, and Dev Bukhsh Singh. 2024. "Biological Big-Data Sources, Problems of Storage, Computational Issues, and Applications: A Comprehensive Review." *Knowledge and Information Systems* 66 (6): 3159–3209.

Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (1): 157.

Emms DM, Kelly S. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Emms, David M., Yi Liu, Laurence J. Belcher, Jonathan Holmes, and Steven Kelly. 2025. "OrthoFinder: Scalable Phylogenetic Orthology Inference for Comparative Genomics." *BioRxiv*. https://doi.org/10.1101/2025.07.15.664860.

Giddey, Alexander D., Elise de Kock, Kehilwe C. Nakedi, Shaun Garnett, Andrew J. M. Nel, Nelson C. Soares, and Jonathan M. Blackburn. 2017. "A Temporal Proteome Dynamics Study Reveals the Molecular Basis of Induced Phenotypic Resistance in Mycobacterium Smegmatis at Sub-Lethal Rifampicin Concentrations." *Scientific Reports* 7 (1). https://doi.org/10.1038/srep43858.

Grinberg, M. 2018. *Flask Web Development: Developing Web Applications Python*. O'Reilly Media.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. "EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Huntley, Rachael P., Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O'Donovan. 2015. "The GOA Database: Gene Ontology Annotation Updates for 2015." *Nucleic Acids Research* 43 (Database issue): D1057-63.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. "GOATOOLS: A Python Library for Gene Ontology Analyses." *Scientific Reports* 8 (1). https://doi.org/10.1038/s41598-018-28948-z.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. "UpSet: Visualization of Intersecting Sets." *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1983–92.

Linard, Benjamin, Ingo Ebersberger, Shawn E. McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, et al. 2021. "Ten Years of Collaborative Progress in the Quest for Orthologs." *Molecular Biology and Evolution* 38 (8): 3033–45.

Lobb, Briallen, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. 2020. "An Assessment of Genome Annotation Coverage across the Bacterial Tree of Life." *Microbial Genomics* 6 (3). https://doi.org/10.1099/mgen.0.000341.

UniProt Consortium. 2025. "UniProt: The Universal Protein Knowledgebase in 2025." *Nucleic Acids Research* 53 (D1): D609–17.

Warren, Andrew S., Jeremy Archuleta, Wu-Chun Feng, and João Carlos Setubal. 2010. "Missing Genes in the Annotation of Prokaryotic Genomes." *BMC Bioinformatics* 11 (1): 131.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. "Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center." *Nucleic Acids Research* 45 (D1): D535–42.