The Journal of Open Source Software

PPanGGOLiN V2: technical enhancement and extended functionalities for prokaryotic pangenome analysis

Jérôme Arnoux [1*], **Jean Mainguy** [1*], **Adelme Bazin** [2], **Guillaume Gautreau** [3], **Téo Lemane** [1], **David Vallenet** [1], **and Alexandra Calteau** [1¶]

1 LABGeM, Genomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, France. **2** Laboratory of Biology and Modeling of the Cell, Ecole Normale Supérieure de Lyon, CNRS, Université Claude Bernard Lyon, Université de Lyon, France **3** Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.11101

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @hkaspersen
- @ramadatta
- @mjsull

Submitted: 24 June 2026
Published: 14 November 2026

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

The exponential growth of genomic data, particularly for microbes, has made pangenomic approaches a gold standard for large-scale comparative genomics. By capturing the full genomic diversity of a species rather than relying on a single reference, pangenomics has transformed microbial genomics, revealing the adaptive potential of bacteria and the evolutionary dynamics underlying functional diversity. Among available tools, PPanGGOLiN distinguishes itself through its graph-based model coupled with statistical gene partitioning.

Here we present PPanGGOLiN v2, which introduces substantial improvements across three dimensions: new analytical features that expand what users can investigate, a comprehensive software architecture redesign that improves maintainability and extensibility, and performance improvements that address the computational demands of ever-growing genomic datasets.

Statement of need

In the last decade, the number of sequenced microbial genomes has exploded from a few thousand to several millions. While this wealth of data offers immense potential, traditional genome-centric approaches reach their limits at this scale. Pairwise genome comparisons become increasingly computationally demanding as the number of genomes grows, while making it challenging to capture patterns of variation shared across large numbers of genomes (Land et al., 2015; Parks et al., 2018). A pangenome condenses all genomic information from a set of genomes into a single structure, capturing gene presence/absence and genetic context: e.g. mobile genetic elements or functional annotations, at the species or clade level (Tettelin et al., 2005).

PPanGGOLiN (Gautreau et al., 2020) compresses genomic information into a computationally efficient graph-based model. Genes are clustered into families based on sequence similarity, while gene neighborhood relationships are encoded in the graph. This unified structure captures genomic diversity across multiple genomes, making it possible to identify patterns of variation shared among them. Gene families are then partitioned using a Bernoulli Mixture Model coupled with a Markov Random Field accounting for both family occurrence and genomic neighborhood. The software suite further integrates PanRGP (Bazin et al., 2020), for identification of Regions of Genomic Plasticity (RGPs) and their insertion spots, and PanModule (Bazin et al., 2021), for finding conserved modules.

PPanGGOLiN v2 builds on this foundation by adding new analytical features: pangenome projection, genomic context search, RGP clustering, and metadata association (Figure 1) together with a full architectural redesign and a modernized development workflow, making the tool more capable, maintainable, and open to community contributions.

State of the field

Pangenomics has transformed prokaryotic genome analysis by moving beyond a reference-centric approach toward representations that capture the entire genomic diversity of microbial populations. Two pangenome graph models are widely used. Sequence-level graphs, like Bifrost (Holley & Melsted, 2020) or vg toolkit (Garrison et al., 2018), encode nucleotide-level variation, providing small variants resolution, but at high computational and interpretive complexity. Gene-level graphs, such as Panaroo (Tonkin-Hill et al., 2020) or pangene (Li et al., 2024), instead cluster genes into families to form graph nodes, with edges encoding genomic contiguity. This representation neglects fine-scale variation but remains highly informative about gene presence/absence and mobile elements at a fraction of the cost.

Most gene-level tools rely on a predefined occurrence threshold to split gene families into different partitions, leading to inconsistent predictions depending on genome quality and diversity. PPanGGOLiN instead uses a statistical method relying on both gene family occurrence patterns and graph topology to partition families into: *persistent*, *shell*, and *cloud* genomes. Among dozens of gene-level pangenomic tools, only mOTUpan (Buck et al., 2022) and micropan (Snipen & Liland, 2015) also perform statistical partitioning. However, these approaches rely solely on gene family occurrence and do not incorporate gene neighborhood information.

Additionally, PPanGGOLiN builds on its graph-based representation and fine-grained partitioning to identify RGPs, spots and conserved modules; analyses unique to PPanGGOLiN.

New features

`projection`: pangenome annotation of external genomes

The projection feature annotates external genomes using a previously computed pangenome, without recomputing the full structure. Input genes are assigned to existing gene families and partitions based on sequence similarity; unmatched genes are considered genome-specific and assigned to the cloud partition. From these assignments, PPanGGOLiN further predicts RGPs, spots, and modules present in the projected genome. This is useful for comparing newly sequenced genomes against an established reference pangenome.

In Version 2, PPanGGOLiN introduces a Proksee-compatible output format, to visualize the circular genomes annotated with pangenome information, from the pangenome itself or via `projection`, allowing further interactive analyses within Proksee (Grant et al., 2023).

`context`: genomic context extraction

A genomic context refers to a set of genes conserved within similar genomic regions across multiple genomes. The pangenome graph already encodes neighborhood relationships, making it well suited for extracting such context. The complete method is described in detail by Arnoux et al. (2026). This feature provides a species-level view of conserved genomic neighborhoods, letting users identify functionally related genes and, for example, accurately reconstruct metabolic pathways.

`rgp_cluster`: clustering of RGPs

A new feature allows comparative analysis of RGPs across genomes in the pangenome. Clustering is based on gene content similarity, where two RGPs are considered related if

their genes belong to the same families, quantified using a Gene Repertoire Relatedness (GRR) score.

Clustering is performed through graph modeling: each RGP is a node, and an edge is added between two nodes if their GRR (min or max) exceeds a threshold (0.8 by default). The resulting graph is then partitioned into clusters using the Louvain community detection algorithm (Blondel et al., 2008), identifying groups of RGPs with similar gene repertoires. Results are exportable as a tabulated file or in formats compatible with graph visualization tools such as Gephi (Bastian et al., 2009) and Cytoscape (Shannon et al., 2003), facilitating the exploration of the diversity and dissemination of mobile genetic elements across genomes.

`metadata`: associating metadata to pangenome elements

Another key addition is the ability to associate metadata with any pangenome element, including genes, contigs, genomes, gene families, edges, RGPs, spots, and modules. Metadata is provided as a tabulated file requiring at minimum the target element's identifier, with no restriction on the type or number of additional fields, making the format highly flexible. Each entry is linked to a named source, allowing an element to carry multiple annotations from different sources.

For performance and portability, metadata is stored directly in the pangenome file. While not used directly in pangenome computation, it is propagated to all outputs, facilitating downstream interpretation and exploration.

Combined with `rgp_cluster` and `projection`, metadata association bridges the gap between pangenome structure and biological interpretation, letting users connect genomic variation to functional, ecological, or experimental data.

Software design

PPanGGOLiN is built around a central data model in which the HDF5 pangenome file serves as the reference data object for the entire system. Analyses can be performed step by step or through workflow commands executing the pipeline in a single run, providing flexibility and convenience (Figure 1). Once executed, the HDF5 pangenome file supports multiple uses: export to other formats for downstream analyses, input for additional PPanGGOLiN commands, and programmatic access through the Python API. This central-object design also enhances reproducibility: parameters are resolved consistently (command line, then configuration file, then defaults), validated across analysis steps, and stored within the pangenome file, enabling analyses to be reliably reproduced and compared.

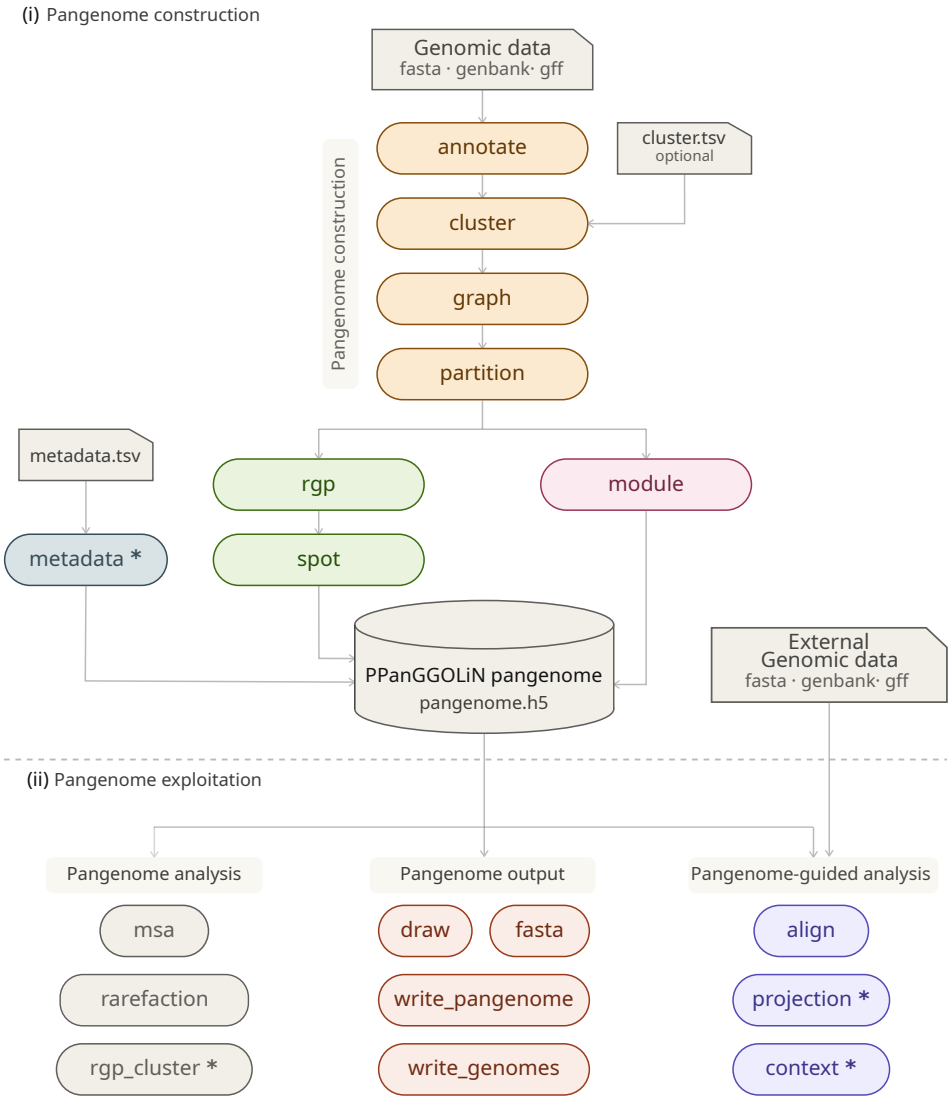


Figure 1: Overview of the PPanGGOLiN v2 workflow. Rounded boxes represent software commands. Commands marked with * are new in v2. Commands fall into two sections. (*i*) *Pangenome construction*: from genomic data (annotated or not), to progressively build the pangenome graph and partition gene families into *persistent*, *shell*, and *cloud*. Two independent analyses then follow: `rgp` and `spot` identify RGPs and their insertion spots, while `module` identifies conserved genomic modules. Results are stored in a single HDF5 file (`pangenome.h5`), the central data object. External metadata can be associated with any pangenome element via `metadata` (*), using a tabulated file. (*ii*) *Pangenome exploitation*: the HDF5 file serves as input for four categories of downstream commands. *Pangenome analysis*: `msa` computes multiple sequence alignments for gene families; `rarefaction` computes the pangenome's rarefaction curve; `rgp_cluster` (*) clusters RGP based on shared gene content. *Pangenome-guided analysis*: `align` maps an external genome or protein set to pangenome families; `context` (*) extracts conserved genomic neighborhoods around genes of interest; `projection` (*) annotates a new genome using an existing pangenome. *Pangenome output*: `draw`, `fasta`, `write_pangenome`, and `write_genomes` export results in various formats for downstream analyses and visualization.

Technical enhancements

PPanGGOLiN was originally developed as a research project. Despite robust and user-friendly workflows, its codebase required a complete redesign to meet long-term software engineering standards.

Several targeted improvements address performance and reproducibility. Prodigal (Hyatt et al., 2010) was replaced by Pyrodigal (Larralde, 2022), reducing I/O overhead during annotation. The HDF5 pangenome file structure was redesigned to drastically reduce file size (Figure 2) and improve API access. Pangenome file reading was optimized, reducing loading time from minutes to seconds for large datasets. Memory consumption during sequence writing was reduced by reading sequences directly from HDF5 files rather than loading them into memory. A configuration file system now facilitates PPanGGOLiN integration into computational platforms (see Research impact).

Finally, PPanGGOLiN v2 is distributed via Bioconda (Grüning et al., 2018) and PyPI, ensuring broad accessibility across computing environments. Its documentation was fully revised and is now hosted on ReadTheDocs: ppanggolin.readthedocs.io, lowering the barrier for new users.

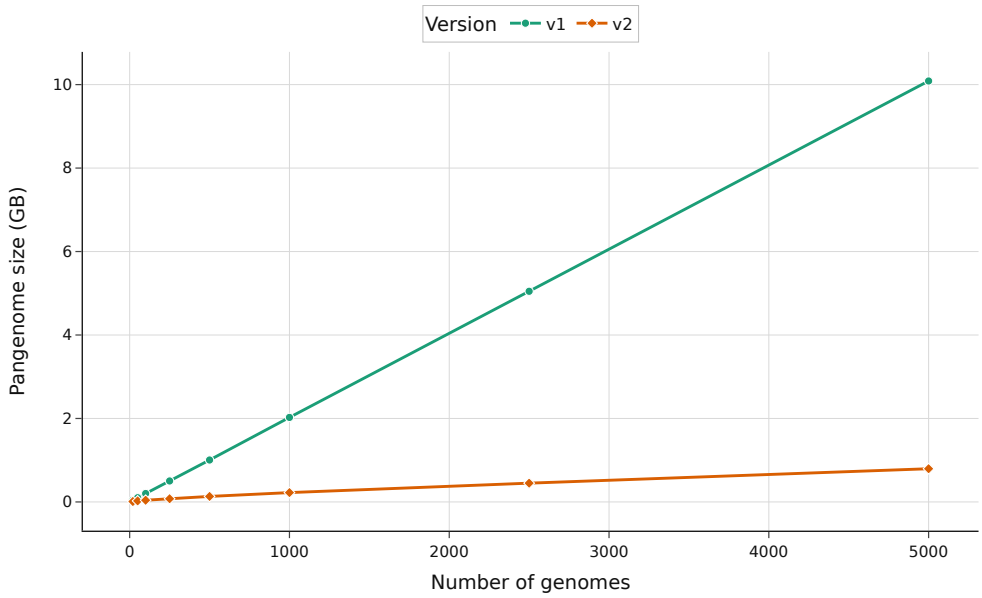


Figure 2: Pangenome HDF5 file size (GB) generated by PPanGGOLiN v1.2.74 vs v2.3.0 across increasing genome counts. PPanGGOLiN v2 consistently produces smaller files, supporting the impact of the v2 file structure redesign.

Research impact statement

PPanGGOLiN is widely adopted in microbial genomics, with ~300 citations on Google Scholar. Its results are integrated into MicroScope (Vallenet et al., 2020), a microbial genome annotation and analysis platform, and a dedicated Galaxy (The Galaxy Community, 2024) module broadens accessibility. The tool is a core dependency for PANORAMA (Arnoux et al., 2026), enabling cross-species biological system predictions and pangenome comparisons, and supports PanGBank (Mainguy et al., 2026), a pangenome database. A companion project further extends this ecosystem by integrating PPanGGOLiN pangenomes — including RGP, spots, and modules — alongside functional metadata into a Neo4j graph database, enabling advanced

graph-based queries across pangenome elements (Arnoux et al., 2024). PPanGGOLiN is actively maintained and evolves through community contributions, bug reports, and feature suggestions. In 2023, it received the Open Science Award for Open Source Research Software from the French Ministry of Higher Education and Research.

AI usage disclosure

AI tools were used in a limited capacity during later development. GitHub Copilot (mainly Claude Sonnet 4.6) assisted with code review suggestions, test writing, and documentation. Claude Sonnet 4.6 (Anthropic) assisted in designing and refining the workflow overview figure (Figure 1) and in manuscript proofreading. Core software design, implementation, and architectural decisions were developed independently without AI assistance. All AI-generated outputs were systematically reviewed and validated by authors.

Acknowledgments

This research was supported by the CFR PhD program of the French Alternative Energies and Atomic Energy Commission (CEA) for JA and the BlueRemediomics project for JM, funded by the European Union under the Horizon Europe Program, Grant Agreement No. 101082304. We thank the PPanGGOLiN user community for their contributions, feedback, and bug reports, which have continuously improved the tool and shaped this new version.

References

Arnoux, J., Bonifati, A., Calteau, A., Dumbrava, S., & Gautreau, G. (2024). Integrating complex pangenome graphs. *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, 350–354. https://doi.org/10.1109/ICDEW61823.2024.00052

Arnoux, J., Mainguy, J., Bry, L., Grado, Q. F. de, Hoblos, Y., Vallenet, D., & Calteau, A. (2026). Panorama: A robust pangenome-based method for predicting and comparing biological systems across species. *PLOS Computational Biology*, *22*(7), e1013856. https://doi.org/10.1371/journal.pcbi.1013856

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D., & Calteau, A. (2020). panRGP: A pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, *36*, i651–i658. https://doi.org/10.1093/bioinformatics/btaa792

Bazin, A., Medigue, C., Vallenet, D., & Calteau, A. (2021). *panModule: Detecting conserved modules in the variable regions of a pangenome graph*. bioRxiv. https://doi.org/10.1101/2021.12.06.471380

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Buck, M., Mehrshad, M., & Bertilsson, S. (2022). mOTUpan: A robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genomics and Bioinformatics*, *4*(3), lqac060. https://doi.org/10.1093/nargab/lqac060

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph

toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879. https://doi.org/10.1038/nbt.4227

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C., & Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, *16*(3), e1007732. https://doi.org/10.1371/journal.pcbi.1007732

Grant, J. R., Enns, E., Marinier, E., Mandal, A., Herman, E. K., Chen, C., Graham, M., Van Domselaar, G., & Stothard, P. (2023). Proksee: In-depth characterization and visualization of bacterial genomes. *Nucleic Acids Research*, *51*, W484–W492. https://doi.org/10.1093/nar/gkad326

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Holley, G., & Melsted, P. (2020). Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, *21*(1), 249. https://doi.org/10.1186/s13059-020-02135-8

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., & Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*(2), 141–161. https://doi.org/10.1007/s10142-015-0433-4

Larralde, M. (2022). Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, *7*(72), 4296. https://doi.org/10.21105/joss.04296

Li, H., Marin, M., & Farhat, M. R. (2024). Exploring gene content with pangene graphs. *Bioinformatics*, *40*(7), btae456. https://doi.org/10.1093/bioinformatics/btae456

Mainguy, J., Lemane, T., Bazin, A., Arnoux, J., Gautreau, G., Médigue, C., Calteau, A., & Vallenet, D. (2026). PanGBank: A large-scale resource of precomputed microbial pangenomes built with PPanGGOLiN. *Nucleic Acids Research*, gkag948. https://doi.org/10.1093/nar/gkag948

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004. https://doi.org/10.1038/nbt.4229

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Snipen, L., & Liland, K. H. (2015). Micropan: An R-package for microbial pan-genomics. *BMC Bioinformatics*, *16*(1), 79. https://doi.org/10.1186/s12859-015-0517-0

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proceedings of the*

National Academy of Sciences of the United States of America, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

The Galaxy Community. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*, *52*, W83–W94. https://doi.org/10.1093/nar/gkae410

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 180. https://doi.org/10.1186/s13059-020-02090-4

Vallenet, D., Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., Burlot, L., Bussell, X., Fouteau, S., Gautreau, G., Lajus, A., Langlois, J., Planel, R., Roche, D., Rollin, J., Rouy, Z., Sabatet, V., & Médigue, C. (2020). MicroScope: An integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, *48*, D579–D589. https://doi.org/10.1093/nar/gkz926