The Journal of Open Source Software

# PANKEGG: Integrative Visualisation and Comparison of Metagenome-Assembled Genomes Annotation, Taxonomy, and Quality

**Renaud Van Damme** [1¶], **Arnaud Vanbelle** [1], **Juliette Hayer** [2,3], **Amrei Binzer-Panchal** [1], and **Erik Bongcam-Rudloff** [1]

**1** Department of Animal Biosciences, Faculty of Center for Veterinary Medicine and Animal Science, Swedish University of Agricultural Sciences, Uppsala, Sweden **2** MIVEGEC, University of Montpellier, IRD, CNRS, Montpellier, France **3** Laboratoire Mixte International Drug Resistance in Southeast Asia ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** Open Journals

**Reviewers:**

- @openjournals

**Submitted:** 01 January 1970
**Published:** unpublished

## Summary

Pankegg is an interactive software suite for parsing and visualising metagenome-assembled genomes (MAGs) and exploring their metabolic capabilities. It integrates quality metrics, annotation, and taxonomic classification in one interactive central database. Pankegg enables researchers to explore, compare, and interpret their data through a modern browser-based interface, streamlining the analysis of large and complex metagenomic datasets. The software supports output from widely used tools such as CheckM2 (Chklovski et al., 2022), GTDB-TK (Chaumeil et al., 2020), Sourmash (Brown & Irber, 2016), and EggNOG (Cantalapiedra et al., 2021), making it a flexible solution for a wide range of microbiome and environmental genomics studies. By interconnecting the different analysis results, people can investigate different interactions in the data more visually and conveniently. Pankegg answers questions such as:

- How many of my bins pass the GTDB (Parks et al., 2022) quality threshold?
- What is their taxonomic classification?
- Which KEGG orthologs are present?
- Which proportion of their respective metabolic pathways is covered by the KEGG orthologs identified?

Answering these questions typically requires inspecting multiple result files and cross-referencing the information. In contrast, Pankegg provides an all-in-one platform to explore, visualise, interpret, and save the results.

## Statement of need

The ever-growing progress of sequencing technologies has made it possible to recover thousands of draft and high-quality genomes directly from a plethora of environmental samples, accelerating our understanding of microbial diversity across ecosystems. Shotgun metagenomics with assembly-based approaches recovers metagenome-assembled genomes (MAGs), giving access to taxonomic and functional profiles of uncultured microorganisms.

Pankegg integrates taxonomic analysis with KEGG pathway annotations, distinguishing itself from tools like MAGFlow/BIgMAG (Yepes-García & Falquet, 2024) and Anvi'o (Eren et al., 2021). While MAGFlow/BIgMAG and Anvi'o metagenomics provide comprehensive metagenomic workflows and their visualisation, Pankegg focuses on the visualisation and the functional interpretation of orthologs' variations across multiple samples and bins within the context of KEGG pathways (Kanehisa et al., 2023). Our focused approach on the KEGG orthologs facilitates a more direct and efficient analysis of metabolic capabilities and variations in microbial communities, even with growing datasets.

As the volume and complexity of metagenomic data increase, so do the challenges of efficiently comparing and visualising results from diverse annotation, classification, and quality assessment tools. In just one year (April 2024 to April 2025), over 135,000 new genomes were added to the Genome Taxonomy Database (GTDB). Tools like CheckM2, Sourmash, GTDB-TK, and EggNOG provide key outputs for quality, taxonomy, and functional annotation, but downstream integration and visualisation remain non-trivial.

Pankegg enables users to merge results from any pipeline, workflow, or manual analysis that provides annotation, classification, and quality information into a standardised SQL database. The database allows users to explore the data through an interactive local web application. The tool is designed to analyse finalised metagenome-assembled genomes (MAGs) and critically evaluate bins obtained during the binning stage of assembly-based metagenomic analysis. By integrating CheckM2 quality metrics, annotation, and taxonomic classification, Pankegg helps users determine which bins meet the GTDB standards to be classified and reported as MAGs and which bins should be excluded due to low quality or inconsistency. Pankegg allows the user to explore and compare the metabolic capabilities of microbial communities.

Pankegg relies on widely used coding languages (Python, JavaScript, and HTML), SQLite as the SQL database engine (Hipp, 2000--2024), and libraries:

- flask (Ronacher & contributors, 2024b)
- jinja2 (Ronacher & contributors, 2024c)
- pandas (team, 2024)
- numpy (Harris et al., 2020)
- SciKit-Learn (Pedregosa et al., 2011)
- SciPy (Virtanen et al., 2020)
- click (Ronacher & contributors, 2024a)
- Python SQLite3 (Foundation, 2024)

Making its installation straightforward in most systems through pip (Authority, 2008), conda (Anaconda, 2012), and pixi (Prefix.dev, 2023), see the Pankegg installation chapter in our documentation. The software installation was tested on Ubuntu, WSL (Ubuntu 16 to 22), Windows 10 & 11, MacOS, and HPE Cray EX supercomputer systems. This unified approach reduces the barriers to integrative metagenomic analysis, enabling both specialists and non-specialists to make informed decisions based on large-scale, genome-resolved metagenomic data.

## Tool Overview

Pankegg consists of two primary tools:

### PANGEGG MAKE DB

This script ingests a CSV file specifying the locations of EggNOG annotations, classification (Sourmash or GTDB-TK), and quality metrics (Checkm2) files for each sample. It then constructs an SQL database aggregating all relevant results.

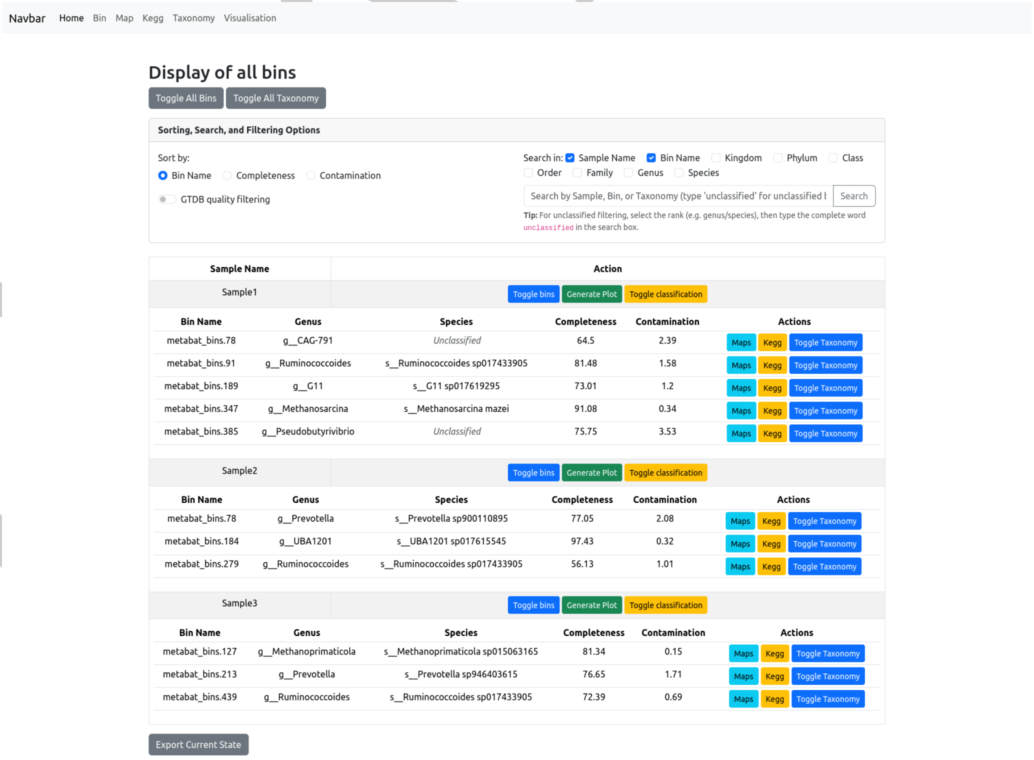
### PANKEGG APP

The app is a Flask-based web server that connects to the SQL database generated by `pankegg_make_db.py`, providing a simple interface for interfacing, cross-referencing, and filtering the data. The interface offers many different information, and each page can be filtered by information from the other pages. A more detailed explanation is in our documentation's Pankegg Web Page chapter. Here is a concise summary:

1. Bin page: outlines all bins/MAGs in the database. Figure 1
2. Pathway page, also called Map page: lists the pathways in the database. Each pathway contains a "completion value". This indicator is calculated by dividing the number of KEGG orthologs in the user's database by the total number of orthologs in the pathway. Figure 2
3. KEGG page: lists the KEGG orthologs in the data. The information for each ortholog is expandable, and the view then includes the corresponding EggNOG entries (bin ID, sample ID, GO terms, KEGG orthologs associated, EggNOG description).
4. Taxonomy page: presents the taxonomic composition of the input database.
5. The sample vs. Sample and Bins vs. Bins pages allow users to compare different samples or bins, respectively, with regard to pathway presence, completeness, and quality metrics.
6. PCA page: This page visualizes a principal component analysis (PCA) based on functional or taxonomic profiles. Beware that PCA interpretation is only valid with enough data; we recommend at least 40 bins [Shaukat2016173190].

Pankegg's app is designed to make exploration intuitive. It features sortable and filterable tables, interactive plots, and external links to KEGG and other databases. These features collectively support users in hypothesis generation, genome curation, and discovering ecological and functional trends in complex datasets.

# Figures



Display of all bins

Sorting, Search, and Filtering Options

Sort by: ● Bin Name ○ Completeness ○ Contamination

GTDB quality filtering

Search in: ☑ Sample Name ☑ Bin Name ☐ Kingdom ☐ Phylum ☐ Class ☐ Order ☐ Family ☐ Genus ☐ Species

Search by Sample, Bin, or Taxonomy (type 'unclassified' for unclassified t

**Tip:** For unclassified filtering, select the rank (e.g. genus/species), then type the complete word `unclassified` in the search box.

| Sample Name | Action |
|---|---|
| Sample1 | Toggle bins Generate Plot Toggle classification |

| Bin Name | Genus | Species | Completeness | Contamination | Actions |
|---|---|---|---|---|---|
| metabat_bins.78 | g__CAG-791 | *Unclassified* | 64.5 | 2.39 | Maps Kegg Toggle Taxonomy |
| metabat_bins.91 | g__Ruminococcoides | s__Ruminococcoides sp017433905 | 81.48 | 1.58 | Maps Kegg Toggle Taxonomy |
| metabat_bins.189 | g__G11 | s__G11 sp017619295 | 73.01 | 1.2 | Maps Kegg Toggle Taxonomy |
| metabat_bins.347 | g__Methanosarcina | s__Methanosarcina mazei | 91.08 | 0.34 | Maps Kegg Toggle Taxonomy |
| metabat_bins.385 | g__Pseudobutyrivibrio | *Unclassified* | 75.75 | 3.53 | Maps Kegg Toggle Taxonomy |

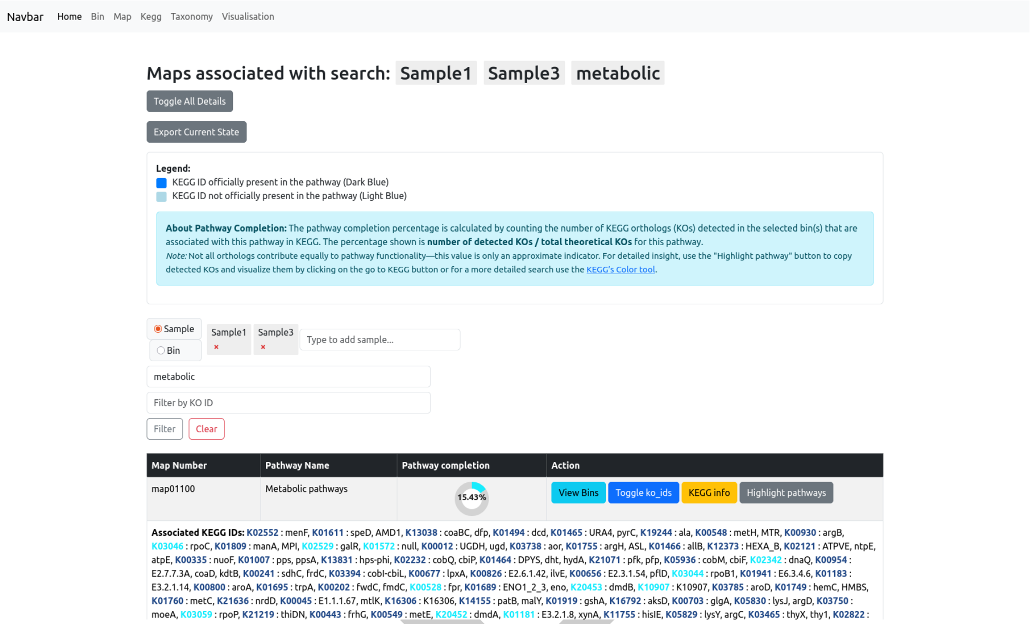
| Sample Name | Action |
|---|---|
| Sample2 | Toggle bins Generate Plot Toggle classification |

| Bin Name | Genus | Species | Completeness | Contamination | Actions |
|---|---|---|---|---|---|
| metabat_bins.78 | g__Prevotella | s__Prevotella sp900110895 | 77.05 | 2.08 | Maps Kegg Toggle Taxonomy |
| metabat_bins.184 | g__UBA1201 | s__UBA1201 sp017615545 | 97.43 | 0.32 | Maps Kegg Toggle Taxonomy |
| metabat_bins.279 | g__Ruminococcoides | s__Ruminococcoides sp017433905 | 56.13 | 1.01 | Maps Kegg Toggle Taxonomy |

| Sample Name | Action |
|---|---|
| Sample3 | Toggle bins Generate Plot Toggle classification |

| Bin Name | Genus | Species | Completeness | Contamination | Actions |
|---|---|---|---|---|---|
| metabat_bins.127 | g__Methanoprimaticola | s__Methanoprimaticola sp015063165 | 81.34 | 0.15 | Maps Kegg Toggle Taxonomy |
| metabat_bins.213 | g__Prevotella | s__Prevotella sp946403615 | 76.65 | 1.71 | Maps Kegg Toggle Taxonomy |
| metabat_bins.439 | g__Ruminococcoides | s__Ruminococcoides sp017433905 | 72.39 | 0.69 | Maps Kegg Toggle Taxonomy |

**Figure 1:** Bin page: This page shows the bins for each sample in the database. Three samples, with eleven bins, are displayed, including their classification, CheckM2 completeness, and contamination.



Maps associated with search: Sample1 Sample3 metabolic

**Legend:**
🟦 KEGG ID officially present in the pathway (Dark Blue)
🟦 KEGG ID not officially present in the pathway (Light Blue)

**About Pathway Completion:** The pathway completion percentage is calculated by counting the number of KEGG orthologs (KOs) detected in the selected bin(s) that are associated with this pathway in KEGG. The percentage shown is **number of detected KOs / total theoretical KOs** for this pathway.
*Note:* Not all orthologs contribute equally to pathway functionality—this value is only an approximate indicator. For detailed insight, use the "Highlight pathway" button to copy detected KOs and visualize them by clicking on the go to KEGG button or for a more detailed search use the KEGG's Color tool.

| Map Number | Pathway Name | Pathway completion | Action |
|---|---|---|---|
| map01100 | Metabolic pathways | 15.43% | View Bins Toggle ko_ids KEGG info Highlight pathways |

**Associated KEGG IDs:** K02552 : menF, K01611 : speD, AMD1, K13038 : coaBC, dfp, K01494 : dcd, K01465 : URA4, pyrC, K19244 : ala, K00548 : metH, MTR, K00930 : argB, K03046 : rpoC, K01809 : manA, MPI, K02529 : galR, K01572 : null, K00012 : UGDH, ugd, K03738 : aor, K01755 : argH, ASL, K01466 : allB, K12373 : HEXA_B, K02121 : ATPVE, ntpE, atpE, K00335 : nuoF, K01007 : pps, ppsA, K13831 : hps-phi, K02232 : cobQ, cbiP, K01464 : DPYS, dht, hydA, K21071 : pfk, pfp, K05936 : cobM, cbiF, K02342 : dnaQ, K00954 : E2.7.7.3A, coaD, kdtB, K00241 : sdhC, frdC, K03394 : cobI-cbiL, K00677 : lpxA, K00826 : E2.6.1.42, ilvE, K00656 : E2.3.1.54, pflD, K03044 : rpoB1, K01941 : E6.3.4.6, K01183 : E3.2.1.14, K00800 : aroA, K01695 : trpA, K00202 : fwdC, fmdC, K00528 : fpr, K01689 : ENO1_2_3, eno, K20453 : dmdB, K10907 : K10907, K03785 : aroD, K01749 : hemC, HMBS, K01760 : metC, K21636 : nrdD, K00045 : E1.1.1.67, mtlK, K16306 : K16306, K14155 : patB, malY, K01919 : gshA, K16792 : aksD, K00703 : glgA, K05830 : lysJ, argD, K03750 : moeA, K03059 : rpoP, K21219 : thiDN, K00443 : frhG, K00549 : metE, K20452 : dmdA, K01181 : E3.2.1.8, xynA, K11755 : hisIE, K05829 : lysY, argC, K03465 : thyX, thy1, K02822 :

**Figure 2:** Map page, On this page, we see a list of maps from Samples 1 and 3 filtered for pathways containing the word 'metabolic.' Only one pathway is visible here, which is 15.43% complete. Below, a list of KEGG orthologs detected in the samples for this pathway is displayed.

# Author contribution

**Renaud Van Damme:** Conceptualisation and software development (initial version), supervision, software co-development (current version), testing, writing original draft, review, and editing.

**Arnaud Vanbelle:** Conceptualisation, software co-development (current version), testing, review, and editing.

**Juliette Hayer:** Methodology, advising on software development, testing, review, and editing.

**Amrei Binzer-Panchal:** Methodology, advising on software development, testing, review, and editing.

**Erik Bongcam-Rudloff:** Supervision, project administration, advising on software development, testing, review, and editing.

# Acknowledgements

Pankegg was initially conceived as a segment of the MUFFIN pipeline (Damme et al., 2021) for metagenome-assembled genome analysis, and we thank the community for feedback and feature suggestions that made it a stand-alone tool.

We acknowledge the Swedish University of Agricultural Sciences Bioinformatics Infrastructure (SLUBI) for providing HPC resources to help develop Pankegg.

We acknowledge the "Studying abroad Erasmus+" programme of the European Union, which allowed Arnaud Vanbelle to exchange between the Haute École en Hainaut (HEH), Belgium, and the Swedish University of Agricultural Sciences for his Master's thesis. The European Commission's support does not endorse the publication's contents, which reflect the author's views.

## References

Anaconda, Inc. (2012). *Conda: Package, dependency and environment management for any language*. https://conda.io/. https://conda.io/

Authority, P. P. (2008). *Pip - the python package installer*. https://pip.pypa.io/. https://pip.pypa.io/

Brown, C. T., & Irber, L. (2016). Sourmash: A library for MinHash sketching of DNA. *The Journal of Open Source Software*, *1*(5), 27. https://doi.org/10.21105/joss.00027

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chklovski, A., Parks, D. H., Woodcroft, B. J., Tyson, G. W., & Hugenholtz, P. (2022). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Genome Biology*, *23*(1), 260. https://doi.org/10.1186/s13059-022-02809-9

Damme, R. V., Hölzer, M., Viehweger, A., Müller, B., Bongcam-Rudloff, E., & Brandt, C. (2021). Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLOS Computational Biology*, *17*(2), e1008716. https://doi.org/10.1371/journal.pcbi.1008716

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, O. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., … Willis, A. D. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology*, *6*, 3–6. https://doi.org/10.1038/s41564-020-00834-3

Foundation, P. S. (2024). *sqlite3 — DB-API 2.0 interface for SQLite databases*.

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., & others. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hipp, D. R. (2000--2024). *SQLite*. SQLite Consortium.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Morishima, K. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, *51*(D1), D587–D592. https://doi.org/10.1093/nar/gkac963

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org/

Prefix.dev. (2023). *Pixi: A modern package and environment manager for python and conda*. https://prefix.dev/docs/pixi/. https://prefix.dev/docs/pixi/

Ronacher, A., & contributors. (2024a). *Click (version 8.2.1)*. https://palletsprojects.com/p/

click/.

Ronacher, A., & contributors. (2024b). *Flask (version 3.1.1)*. https://flask.palletsprojects.com/.

Ronacher, A., & contributors. (2024c). *Jinja2 (version 3.1.6)*. https://palletsprojects.com/p/jinja/.

team, T. pandas development. (2024). Pandas-dev/pandas: Pandas (version 2.2.3). *Zenodo*. https://doi.org/10.5281/zenodo.3509134

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Walt, S. J. van der, Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Contributors, S. 1.0. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Yepes-García, J., & Falquet, L. (2024). Metagenome quality metrics and taxonomical annotation visualization through the integration of MAGFlow and BIgMAG. *F1000Research*, *13*, 640. https://doi.org/10.12688/f1000research.152290.2