





TAXPASTA: TAXonomic Profile Aggregation and STAndardisation

Moritz E. Beber¹, Maxime Borry^{2,3}, Sofia Stamouli⁴, and James A. Fellows Yates^{2,3,5}

1 Unseen Bio ApS, Copenhagen, Denmark **2** Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany **3** Associated Research Group of Archaeogenetics, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Jena, Germany **4** Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Solna, Sweden **5** Department of Paleobiotechnology, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Jena, Germany ¶ Corresponding author

DOI: [10.21105/joss.05627](https://doi.org/10.21105/joss.05627)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Kevin M. Moerman](#)  

Reviewers:

- [@Kevin-Mattheus-Moerman](#)

Submitted: 05 July 2023

Published: 11 July 2023

License

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

Summary

Metagenomic analysis is largely concerned with untargeted genetic characterisation of the taxonomic and functional composition of whole communities of organisms. Researchers ask questions from metagenomic sequencing such as ‘who is present’ (what organisms are present), and ‘what are they doing’ (which functions are they performing)? The nature of this field is such that it intersects with ecology, medicine, statistics, and bioinformatics. Facilitated by the development of Next-Generation Sequencing (NGS), the field often generates large datasets consisting of many samples (hundreds) and many sequencing reads (tens of millions).

In part, due to the interdisciplinary nature of the field, but more importantly, due to the lack of a gold standard, the task of accurately identifying the taxonomic origin of each sequencing read is a popular and unresolved bioinformatics problem. Furthermore, the sizes of the datasets present interesting challenges for computational efficiency, which may require trading off accuracy for speed and memory use. Thus, there exists a diverse number of bioinformatics tools in order to analyse metagenomic sequencing data and produce metagenomic profiles. However, most of those tools have invented their own (often tabular) result formats, which complicates downstream analysis and in particular comparison across tools.

TAXPASTA is a standalone command-line tool written in Python ([Van Rossum & Drake Jr, 1995](#)) that aims to standardise the diverse range of metagenomic profiler output formats to simple tabular formats that are readily consumed in downstream applications. TAXPASTA facilitates cross-comparison between taxonomic profiling tools without the need for external or dedicated modules or plugins needed of other ‘dedicated’ metagenomic profile formats.

Statement of need

TAXPASTA is a Python package for standardising and aggregating metagenomic profiles coming from a wide range of tools and databases ([Figure 1](#)). It was developed as part of the `nf-core/taxprofiler` pipeline¹ within the `nf-core` community ([Ewels et al., 2020](#)).

Across profilers, relative abundances can be reported in read counts, fractions, or percentages, as well as any number of additional columns with extra information. Taxa can be recorded using taxonomic identifiers, taxonomic names and/or in some cases semi-colon-separated taxonomic ‘paths’ (lineages). These can also be formatted in different ways, from typical

¹See [nf-core/taxprofiler](#) and at DOI [10.5281/zenodo.7728364](https://doi.org/10.5281/zenodo.7728364).

tables, to including ‘indented’ taxonomy trees such as in the Kraken (Wood et al., 2019) family of profilers. Manually parsing these for comparison can be an arduous, error-prone task, with researchers often reverting to custom R (R Core Team, 2023) and Python scripting, or even manual correction in spreadsheet software.

With TAXPASTA, all of those formats can be converted into a single, standardised output, that, at a minimum, contains taxonomic identifiers and their relative abundances as integer counts. It can also be used to aggregate profiles across samples from the same profiler and merge them into a single, standardised table. Having a singular format facilitates downstream analyses and comparisons. TAXPASTA is not the first tool to attempt standardising metagenomic profiles, but it is by far the most comprehensive in terms of supported profilers and output formats.

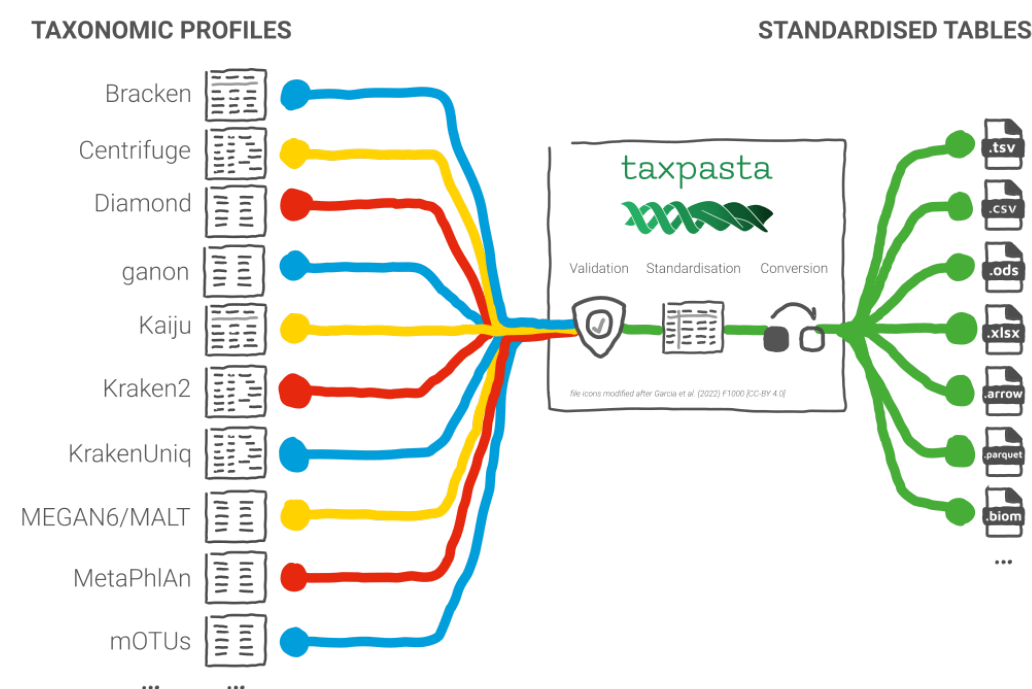


Figure 1: A visual summary of TAXPASTA's capabilities. Supported profilers are listed on the left and output formats on the right.

There exists an initiative to benchmark and compare profilers, as well as provide guidance on their fitness for purpose; the Critical Assessment of Metagenome Interpretation (CAMI) challenges (Meyer et al., 2022; Sczyrba et al., 2017). For that initiative, the Open-community Profiling Assessment tool (OPAL) (Meyer et al., 2019) was developed. Creating a community wide assessment faced many of the challenges presented here, however, the chosen solution was to mandate a single output format² for all profilers participating in the challenge. Furthermore, OPAL is an integrated tool performing assessment and visualisation, whereas TAXPASTA follows the UNIX philosophy³ of doing one thing and doing it well. The BIOM format (McDonald et al., 2012) was created with a similar intention of standardising a storage format for microbiome analyses. However, transforming metagenomic profiles into that format is entirely left up to the user. The format also is not easily loadable into spreadsheet software, and external libraries are required for loading the format into data analysis languages such as R. The QIIMETM2 *next-generation microbiome bioinformatics platform* (Bolyen et al., 2019) also maintains internally consistent formats for storing and processing metagenomic data that

²<https://github.com/bioboxes/rfc/tree/master/data-format>

³https://en.wikipedia.org/wiki/Unix_philosophy#Origin

new tools can plug into, however this suite of software was originally designed for the analysis of 16S rRNA amplicon sequencing data (Caporaso et al., 2010), and whole-genome, shotgun metagenomic sequencing data is only supported via community plugins⁴. While some of the taxonomic profilers also come with scripts to convert their output into another format and/or merge multiple profiles into a single table, such as the Kraken tools companion package (Lu et al., 2022), these are often focused on the specific tool or family of tools. Thus, users would have to become proficient in yet another piece of software per tool or family of tools for the sake of consistent output files.

TAXPASTA supports reading a wide range of formats of primarily shotgun-metagenomic profiling tools and formats, and it is designed to be used as a building block in metagenomic analysis workflows. At the time of writing, it is able to read profiles from nine different profilers, namely Bracken (Lu et al., 2017), Centrifuge (Kim et al., 2016), DIAMOND (Buchfink et al., 2021), ganon (Piro et al., 2020), Kaiju (Menzel et al., 2016), Kraken2 (Wood et al., 2019), KrakenUniq (Breitwieser et al., 2018), MALT/MEGAN6 (Huson et al., 2016; Vågene et al., 2018), MetaPhlAn (Blanco-Míguez et al., 2023), and mOTUs (Ruscheweyh et al., 2022). Supporting more profilers is already planned, and detailed documentation for community contributions is provided⁵.

For maximum compatibility, TAXPASTA offers a wide range of familiar output file formats, such as text-based, tabular formats (CSV⁶, TSV⁷), spreadsheets (ODS⁸, XLSX⁹), optimised binary formats (Apache Arrow¹⁰ and Parquet¹¹), and the HDF5-based¹² BIOM format (McDonald et al., 2012). We hope that this will let researchers plug and play TAXPASTA into their existing analysis workflows in a wide range of settings.

Acknowledgements

SS was supported by “Rapid establishment of comprehensive laboratory pandemic preparedness – RAPID-SEQ” (awarded to Prof. Jan Albert). MB and JAFY were supported by the Max Planck Society. JAFY received funding from the Werner Siemens-Stiftung (“Paleobiotechnology”, awarded to Prof. Pierre Stallforth and Prof. Christina Warinner). MB received funding from the Balance of Microverse Cluster of Excellence (EXC 2051 – project ID 390713860, awarded to Prof. Christina Warinner).

References

- Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., ... Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-023-01688-w>
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and

⁴<https://library.qiime2.org/plugins/>

⁵https://taxpasta.readthedocs.io/en/latest/contributing/supporting_new_profiler/

⁶https://en.wikipedia.org/wiki/Comma-separated_values

⁷https://en.wikipedia.org/wiki/Tab-separated_values

⁸<https://en.wikipedia.org/wiki/OpenDocument>

⁹https://en.wikipedia.org/wiki/Office_Open_XML

¹⁰<https://arrow.apache.org/>

¹¹<https://parquet.apache.org/>

¹²<https://www.hdfgroup.org/solutions/hdf5/>

- extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8), 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Breitwieser, F. P., Baker, D. N., & Salzberg, S. L. (2018). KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1), 198. <https://doi.org/10.1186/s13059-018-1568-0>
- Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368. <https://doi.org/10.1038/s41592-021-01101-x>
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335–336. <https://doi.org/10.1038/nmeth.f.303>
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), 276–278. <https://doi.org/10.1038/s41587-020-0439-x>
- Huson, D. H., Beier, S., Flade, I., Górski, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., & Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Computational Biology*, 12(6), e1004957. <https://doi.org/10.1371/journal.pcbi.1004957>
- Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12), 1721–1729. <https://doi.org/10.1101/gr.210641.116>
- Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3, e104. <https://doi.org/10.7717/peerj-cs.104>
- Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, 17(12), 2815–2839. <https://doi.org/10.1038/s41596-022-00738-y>
- McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., & Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1), 2047–217X-1-7. <https://doi.org/10.1186/2047-217X-1-7>
- Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1), 11257. <https://doi.org/10.1038/ncomms11257>
- Meyer, F., Bremges, A., Belmann, P., Janssen, S., McHardy, A. C., & Koslicki, D. (2019). Assessing taxonomic metagenome profilers with OPAL. *Genome Biology*, 20(1), 51. <https://doi.org/10.1186/s13059-019-1646-y>
- Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., ... McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: The second round of challenges. *Nature Methods*, 19(4), 429–440. <https://doi.org/10.1038/s41592-022-01431-4>
- Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K., & Renard, B. Y. (2020). Ganon: Precise metagenomics classification against large and up-to-date sets of reference sequences.

- Bioinformatics*, 36(Supplement_1), i12–i20. <https://doi.org/10.1093/bioinformatics/btaa458>
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Ruscheweyh, H.-J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M. I., Wirbel, J., Bork, P., Mende, D. R., Zeller, G., & Sunagawa, S. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome*, 10(1), 212. <https://doi.org/10.1186/s40168-022-01410-z>
- Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., ... McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11), 1063–1071. <https://doi.org/10.1038/nmeth.4458>
- Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., Spyrou, M. A., Andrades Valtueña, A., Huson, D., Tuross, N., Bos, K. I., & Krause, J. (2018). *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution*, 2(3), 520–528. <https://doi.org/10.1038/s41559-017-0446-6>
- Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 257. <https://doi.org/10.1186/s13059-019-1891-0>