# sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets

Luiz Irber [1*], **N. Tessa Pierce-Ward** [1*], **Mohamed Abuelanin** [1], **Harriet Alexander** [2], **Abhishek Anant** [9], **Keya Barve** [1], **Colton Baumler** [1], **Olga Botvinnik** [3], **Phillip Brooks** [1], **Daniel Dsouza** [9], **Laurent Gautier** [9], **Mahmudur Rahman Hera** [4], **Hannah Eve Houts** [1], **Lisa K. Johnson** [1], **Fabian Klötzl** [5], **David Koslicki** [4], **Marisa Lim** [1], **Ricky Lim** [9], **Bradley Nelson** [9], **Ivan Ogasawara** [9], **Taylor Reiter** [1], **Camille Scott** [1], **Andreas Sjödin** [6], **Daniel Standage** [7], **S. Joshua Swamidass** [8], **Connor Tiffany** [9], **Pranathi Vemuri** [3], **Erik Young** [1], and **C. Titus Brown** [1¶]

1 University of California Davis, Davis, CA, United States of America **2** Woods Hole Oceanic Institution, Woods Hole, MA, Unites States of America **3** Chan-Zuckerberg Biohub, San Francisco, CA, United States of America **4** Pennsylvania State University, University Park, PA, United States of America **5** Max Planck Institute for Evolutionary Biology, Plön, Germany **6** Swedish Defence Research Agency (FOI), Stockholm, Sweden **7** National Biofrensic Analysis Center, Fort Detrick, MD, United States of America **8** Washington University in St Louis, St Louis, MO, United States of America **9** No affiliation ¶ Corresponding author * These authors contributed equally.

Summary

sourmash is a command line tool and Python library for sketching collections of DNA, RNA, and amino acid k-mers for biological sequence search, comparison, and analysis (Pierce et al., 2019). sourmash's FracMinHash sketching supports fast and accurate sequence comparisons between datasets of different sizes (Irber, Brooks, et al., 2022), including taxonomic profiling (Portik et al., 2022), functional profiling (Rahman Hera, Liu, et al., 2023), and petabase-scale sequence search (Irber, Pierce-Ward, et al., 2022). From release 4.x, sourmash is built on top of Rust and provides an experimental Rust interface.

FracMinHash sketching is a lossy compression approach that represents data sets using a "fractional" sketch containing $1/S$ of the original k-mers. Like other sequence sketching techniques (e.g. MinHash, (Ondov et al., 2015)), FracMinHash provides a lightweight way to store representations of large DNA or RNA sequence collections for comparison and search. Sketches can be used to identify samples, find similar samples, identify data sets with shared sequences, and build phylogenetic trees. FracMinHash sketching supports estimation of overlap, bidirectional containment, and Jaccard similarity between data sets and is accurate even for data sets of very different sizes.

Since sourmash v1 was released in 2016 (Brown & Irber, 2016), sourmash has expanded to support new database types and many more command line functions. In particular, sourmash now has robust support for both Jaccard similarity and Containment calculations, which enables analysis and comparison of data sets of different sizes, including large metagenomic samples. As of v4.4, sourmash can convert these to estimated Average Nucleotide Identity (ANI) values, which can provide improved biological context to sketch comparisons (Rahman Hera, Pierce-Ward, et al., 2023).

Statement of Need

Large collections of genomes, transcriptomes, and raw sequencing data sets are readily available in biology, and the field needs lightweight computational methods for searching and summarizing the content of both public and private collections. sourmash provides a flexible set of programmatic tools for this purpose, together with a robust and well-tested command-line interface. It has been used in over 350 publications (based on citations of Brown & Irber (2016) and Pierce et al. (2019)) and it continues to expand in functionality.

Acknowledgements

References

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

Irber, L. C., Brooks, P. T., Reiter, T. E., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., & Brown, C. T. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *bioRxiv*. https://doi.org/10.1101/2022.01.11.475838

Irber, L. C., Pierce-Ward, N. T., & Brown, C. T. (2022). Sourmash branchwater enables lightweight petabyte-scale sequence search. *bioRxiv*. https://doi.org/10.1101/2022.11.02.514947

Ondov, B. D., Treangen, T. J., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2015). Fast genome and metagenome distance estimation using MinHash. *bioRxiv*, 029827. https://doi.org/10.1101/029827

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019). Large-scale sequence comparisons with sourmash. *F1000Research*, *8*, 1006. https://doi.org/10.12688/f1000research.19675.1

Portik, D. M., Brown, C. T., & Pierce-Ward, N. T. (2022). Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets. *Bioinformatics*. https://doi.org/10.1186/s12859-022-05103-0

Rahman Hera, M., Liu, S., Wei, W., Rodriguez, J. S., Ma, C., & Koslicki, D. (2023). Fast, lightweight, and accurate metagenomic functional profiling using FracMinHash sketches. *bioRxiv*, 2023–2011. https://doi.org/10.1101/2023.11.06.565843

Rahman Hera, M., Pierce-Ward, N. T., & Koslicki, D. (2023). Deriving confidence intervals for mutation rates across a wide range of evolutionary distances using FracMinHash. *Genome Research*, gr–277651. https://doi.org/10.1101/gr.277651.123