The Journal of Open Source Software

sourmash: a tool 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** [5], **Fabian Klötzl** [6], **David Koslicki** [4], **Marisa Lim** [5], **Ricky Lim** [9], **Ivan Ogasawara** [9], **Taylor Reiter** [1], **Camille Scott** [1], **Andreas Sjödin** [7], **Daniel Standage** [8], **S. Joshua Swamidass** [9], **Connor Tiffany** [9], **Pranathi Vemuri** [3], **Erik Young** [1], **and C. Titus Brown** [1¶]

1 University of California, Davis **2** Woods Hole Oceanographic Institute **3** Chan-Zuckerberg Biohub **4** Pennsylvania State University **5** 10x Genomics **6** MPI for Evolutionary Biology **7** Swedish Defence Research Agency (FOI) **8** National Bioforensic Analysis Center **9** No affiliation ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

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 petabase-scale database 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 (Hera et al., 2022).

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 functionality for this purpose, together with a robust and well-tested command-line interface. It has been used in well over 200 publications (based on citations of Brown & Irber (2016) and Pierce et al. (2019)) and it continues to expand in functionality.

Acknowledgements

This work is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative [GBMF4551 to CTB].

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

Hera, M. R., Pierce-Ward, N. T., & Koslicki, D. (2022). Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances. *bioRxiv*.

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*.

Irber, L. C., Pierce-Ward, N. T., & Brown, C. T. (2022). Sourmash branchwater enables lightweight petabyte-scale sequence search. *bioRxiv*.

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