

sourmash: a library for MinHash sketching of DNA

C. Titus Brown¹ and Luiz Irber¹

DOI: [10.21105/joss.00027](https://doi.org/10.21105/joss.00027)

¹ University of California, Davis

Software

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

Licence

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

Summary

sourmash is a toolbox for creating, comparing, and manipulating MinHash sketches of genomic data.

MinHash sketches provide a lightweight way to store “signatures” of large DNA or RNA sequence collections, and then compare or search them using a Jaccard index. MinHash sketches can be used to identify samples, find similar samples, identify data sets with shared sequences, and build phylogenetic trees (Ondov et al. 2015).

sourmash provides a command line script, a Python library, and a CPython module for MinHash sketches.

References

Ondov, Brian D, Todd J Treangen, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. 2015. “Fast Genome and Metagenome Distance Estimation Using Minhash.” *bioRxiv*. Cold Spring Harbor Labs Journals, 029827. doi:10.1101/029827.