The Journal of Open Source Software

# Blini: lightweight nucleotide sequence search and dereplication

**Amit Lavon** [1]

**1** University of California, Irvine, CA, USA

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:**

**Submitted:** 25 June 2025
**Published:** unpublished

## Summary

Blini is a tool for quick lookup of nucleotide sequences in databases, and for quick dereplication of sequence collections. It is meant to help cleaning and characterizing large collections of sequences that would otherwise be too big to search with BLAST (Altschul et al., 1990) or too demanding for a local machine to process, for example with Sourmash (Brown & Irber, 2016) or with MMseqs (Steinegger & Söding, 2018). Blini is designed to be fast and have a small memory footprint, while allowing the user to tweak its resource consumption to improve matching resolution. Finally, Blini is delivered as a single runnable binary, with no need to install any additional software.

## Statement of need

Metagenomes are collections of genetic material from various organisms, which are often not initially known. Characterizing the taxonomic makeup of a sample involves searching its contents in large databases in order to find which organism matches each nucleotide sequence. Assembled sequences can reach lengths of millions of bases, making alignment-based search methods too cumbersome. Such big queries are often outsourced to powerful cloud-based services such as BLAST (Altschul et al., 1990). In recent years, k-mer-based algorithms were introduced, that enabled efficient searching in large datasets on local machines. Mash distance (Ondov et al., 2016) introduced an alignment-free estimation formula for average nucleotide identity between sequences, making sequence comparison linear. Sourmash (Brown & Irber, 2016) uses fractional min-hashing in order to create small representations of large sequences, which allow for efficient searching and comparison. The LinClust clustering algorithm (Steinegger & Söding, 2018) uses k-mer matching reduce the number of pairwise comparisons and achieve linear scaling with the size of the input.

Blini combines insights from Mash, Sourmash, and LinClust into a simple tool that can quickly cluster or look up big collections of sequences using estimated identity or containment, with tweakable estimation resolution (similar to Sourmash's *scale*).

## Performance comparison

### Search

The search function was tested on RefSeq's viral reference (Pruitt et al., 2007). 100 viral genomes were randomly selected for the test. The algorithms were then run on the 100 genomes as queries, and the original database as reference. Each algorithm was expected to match each genome with its source in the database. In a second run, random SNPs were introduced to 1% of the genomes' bases, and the same test was rerun. For each test, the number of matches with sequences other than the query's source was also recorded.

| Test | Time (s) | Successful matches | Non-source matches |
|---|---|---|---|
| Blini | 0.4 | 98 | 254 |
| Blini (SNPs) | 0.4 | 98 | 207 |
| Sourmash | 681 | 97 | 244 |
| Sourmash (SNPs) | 680 | 94 | 139 |

## Clustering

The clustering function was tested on two simulated datasets. In one dataset each sequence had two counterparts with random SNPs. In the second dataset random fragments were extracted from each root sequence. The algorithms were expected to group each sequence with its mutated counterparts or with its fragments. Performance was evaluated using the Adjusted Rand Index (ARI).

Blini's *scale* refers to the fraction of k-mers considered for the operation. Scale 50 means that 1/50 of k-mers were used.

### SNPs dataset

| Test | Time (s) | Max memory (MB) | ARI |
|---|---|---|---|
| MMseqs (1 thread) | 166 | 977 | 1 |
| MMseqs (4 threads) | 54 | 1024 | 1 |
| Blini (scale 25) | 10.9 | 148 | 1 |
| Blini (scale 50) | 9.7 | 78 | 1 |
| Blini (scale 100) | 8.7 | 30 | 1 |
| Blini (scale 200) | 9.2 | 20 | 0.998 |

### Fragments dataset

| Test | Time (s) | Max memory (MB) | ARI |
|---|---|---|---|
| MMseqs (1 thread) | 130 | 181 | 1 |
| MMseqs (4 threads) | 43 | 727 | 1 |
| Blini (scale 25) | 6.7 | 58 | 1 |
| Blini (scale 50) | 6.1 | 33 | 0.999 |
| Blini (scale 100) | 5.9 | 31 | 0.993 |
| Blini (scale 200) | 5.8 | 14 | 0.972 |

## Limitations

Blini is designed to work on sequences ~10 times longer than the selected *scale* value. For the default value of 200, sequences shorter than 2000 are likely to be accidentally missed. While the scale can be tweaked, this tool might not be suitable for short reads.

Blini currently only works for nucleotide sequences. Amino acid sequences might be added in the future.

Blini is currently single-threaded. Multithreading can be considered in the future if a concrete need arises.

## References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

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

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*, 1–14.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl_1), D61–D65.

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1), 2542.