

sam2lca: Lowest Common Ancestor for SAM/BAM/CRAM alignment files

Maxime Borry ¹, Alexander Hübner ^{1,2}, and Christina Warinner ^{1,2,3}

1 Microbiome Sciences Group, Max Planck Institute for Evolutionary Anthropology, Department of Archaeogenetics, Leipzig, Germany **2** Faculty of Biological Sciences, Friedrich-Schiller Universität Jena, Jena, Germany **3** Department of Anthropology, Harvard University, Cambridge, MA, United States of America

DOI: [10.21105/joss.04360](https://doi.org/10.21105/joss.04360)

Software

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

Editor: [Jacob Schreiber](#)  

Reviewers:

- [@fasnicar](#)
- [@marouenbg](#)

Submitted: 21 April 2022

Published: 01 June 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

sam2lca is a program performing reference sequence disambiguation for reads mapping to multiple reference sequences in a shotgun metagenomics sequencing dataset. To do so, it takes as input the common SAM sequence alignment format and applies the lowest common ancestor algorithm.

Statement of need

The rapidly decreasing cost of massively parallel short-read DNA sequencing technologies has enabled the genetic characterization of entire ecological communities, a technique known as shotgun metagenomics.

In a typical shotgun metagenomics approach, after the DNA of an ecological community has been sequenced, it is compared to a genetic reference database of organisms with known taxonomy. Even though the number of DNA sequences and genomes in reference databases is constantly growing, there are still instances where a query sequence will not have a direct match in a reference database, and it will instead weakly align to one or more distantly related reference organisms. Furthermore, when analyzing short DNA sequences, a query DNA sequence will often match equally well to more than one reference organism, posing a challenge for its taxonomic assignment.

One solution to this problem is to apply a lowest common ancestor algorithm (LCA) ([Figure 1](#)) during taxonomic profiling to place such ambiguous assignments higher in a taxonomic tree, where they can be more confidently assigned. This idea was first implemented for metagenomics with the MEGAN program ([Huson et al., 2007](#)).

Many programs have since been developed to perform LCA during taxonomic profiling. For example, MALT ([Herbig et al., 2017](#)) and MetaPhlan ([Segata et al., 2012](#)) perform LCA and taxonomic profiling after DNA sequence alignment, while other programs, such as Kraken2 ([Wood et al., 2019](#)) and Centrifuge ([Kim et al., 2016](#)), are alignment-free methods that apply LCA after k-mer matching. While combining the steps of database matching and LCA into one program can be useful, it also limits user choice for the selection of different alignment or k-mer matching programs.

With sam2lca, we propose to decouple the LCA step from the alignment step to allow the end-user to freely choose from one of the many DNA sequence aligner programs available, such as Bowtie2 ([Langmead & Salzberg, 2012](#)), bwa ([Li & Durbin, 2009](#)), bmap ([Bushnell, 2014](#)), or minimap2 ([Li, 2018](#)). Each of these aligners exports the sequence alignments in

the widely adopted Sequence Alignment Map format (SAM) (Li et al., 2009), or in its binary (BAM), or compressed representation (CRAM), which sam2lca uses as an input.

The use of the SAM file format enables easier integration of sam2lca in a wide variety of analysis workflows, which often already contain steps generating or using SAM/BAM/CRAM files, and allows for an easy subsequent analysis using well-established programs, such as SAMtools (Li et al., 2009).

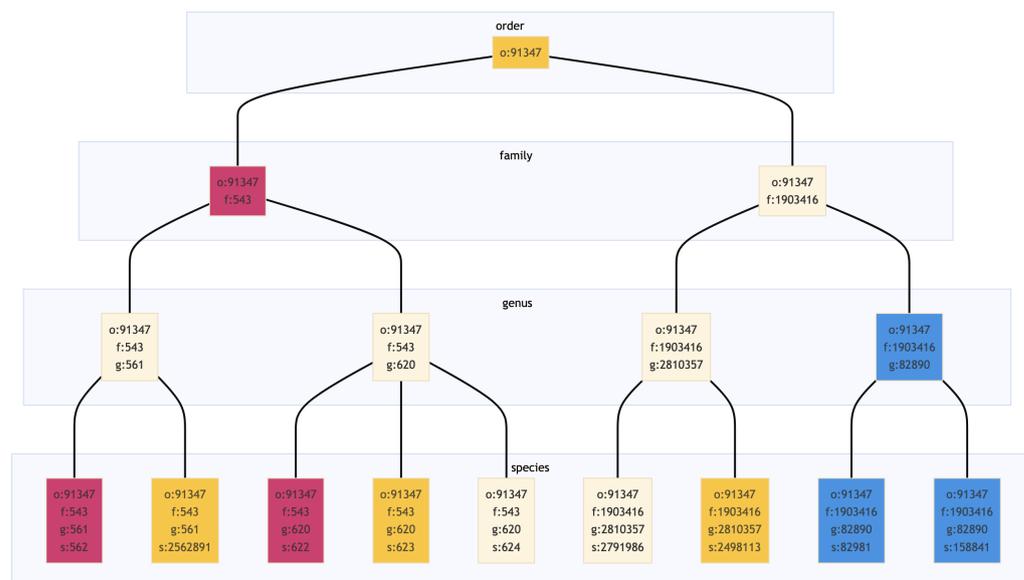


Figure 1: Example of the LCA algorithm with NCBI TAXIDs. Taxons and their LCA are displayed in the same color. The lineage for each taxon is shown with a one letter code for the rank, and the corresponding TAXID. The LCA of s:562 (*E. coli* species) and s:622 (*S. dysenteriae* species) is f:543 (*Enterobacteriaceae* family). The LCA of s:82981 (*L. grimontii* species) and s:158841 (*L. richardii* species) is g:82890 (*Leminorella* genus). The LCA of s:2562891 (*E. alba* species), s:623 (*S. flexneri* species) and s:2498113 (*J. zhutongyuui* species) is o:91347 (Enterobacterales order)

Implementation

sam2lca is a program written in Python, which takes as an input an indexed and sorted SAM/BAM/CRAM alignment file. Broadly, the program consists of four main steps. First, reference sequence accessions, present in the BAM file header section, are converted to taxonomic identifiers (TAXID) using a RocksDB persistent key-value store (Dong et al., 2021). The alignment section of the BAM file is then parsed with Pysam (pysam-developers, 2022) and a dictionary is created to match single and multi-mapping query sequences/reads to the TAXID(s) of their matching reference sequence(s). Next, if a read has been matched to multiple TAXIDs, the LCA implementation of Taxopy (Camargo, 2022) is used to attribute it to the lowest common ancestor, using the NCBI taxonomy by default. Finally, each TAXID is used to retrieve its associated taxon's scientific name and taxonomic lineage, and results are saved in a JSON and CSV file. Optionally, a BAM file, similar to the input file, can be generated. This BAM file contains for each read an additional XT tag added to report the TAXID of the LCA for each read, an XN tag for the taxon's scientific name, and finally an XR tag for the taxon's rank. sam2lca is distributed through pip and conda, and the documentation and tutorials are available at sam2lca.readthedocs.io

Acknowledgements

This research was supported by the Werner Siemens Stiftung (M.B. and C.W.) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2051, Project-ID 390713860 (A.H. and C.W.).

References

- Bushnell, B. (2014). *BBMap: A fast, accurate, splice-aware aligner*. <https://www.osti.gov/biblio/1241166>
- Camargo, A. (2022). Taxopy: A python package for manipulating NCBI-formatted taxonomic databases. In *GitHub repository*. GitHub. <https://github.com/apcamargo/taxopy>
- Dong, S., Kryczka, A., Jin, Y., & Stumm, M. (2021). RocksDB: Evolution of development priorities in a key-value store serving large-scale applications. *ACM Transactions on Storage (TOS)*, 17(4), 1–32. <https://doi.org/10.1145/3483840>
- Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2017). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the tyrolean iceman. *BioRxiv*. <https://doi.org/10.1101/050559>
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386. <https://doi.org/10.1101/gr.5969107>
- 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>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- pysam-developers. (2022). Pysam: A python module for reading and manipulating files in the SAM/BAM format. In *GitHub repository*. GitHub. <https://github.com/pysam-developers/pysam>
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), 811–814. <https://doi.org/10.1038/nmeth.2066>
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/s13059-019-1891-0>