

1 **Struo2: efficient metagenome profiling database construction for ever-expanding**
2 **microbial genome datasets**

3 Nicholas D. Youngblut^{*,1}, Ruth E. Ley¹

4 ¹Department of Microbiome Science, Max Planck Institute for Developmental Biology, Max Planck Ring 5,

5 72076 Tübingen, Germany

6 * Corresponding author: Nicholas Youngblut (nicholas.youngblut@tuebingen.mpg.de)

7 **Running title:** Struo2 builds databases faster

8 **Key words:** metagenome, database, profiling, GTDB

9 Abstract

10 Mapping metagenome reads to reference databases is the standard approach for
11 assessing microbial taxonomic and functional diversity from metagenomic data. However, public
12 reference databases often lack recently generated genomic data such as
13 metagenome-assembled genomes (MAGs), which can limit the sensitivity of read-mapping
14 approaches. We previously developed the Struo pipeline in order to provide a straight-forward
15 method for constructing custom databases; however, the pipeline does not scale well with the
16 ever-increasing number of publicly available microbial genomes. Moreover, the pipeline does
17 not allow for efficient database updating as new data are generated. To address these issues,
18 we developed Struo2, which is >3.5-fold faster than Struo at database generation and can also
19 efficiently update existing databases. We also provide custom Kraken2, Bracken, and
20 HUMAnN3 databases that can be easily updated with new genomes and/or individual gene
21 sequences. Struo2 enables feasible database generation for continually increasing large-scale
22 genomic datasets.

23 Availability:

- 24 • Struo2: <https://github.com/leylabmpi/Struo2>
- 25 • Pre-built databases: <http://ftp.tue.mpg.de/ebio/projects/struo2/>
- 26 • Utility tools: https://github.com/nick-youngblut/gtdb_to_taxdump

27 Results

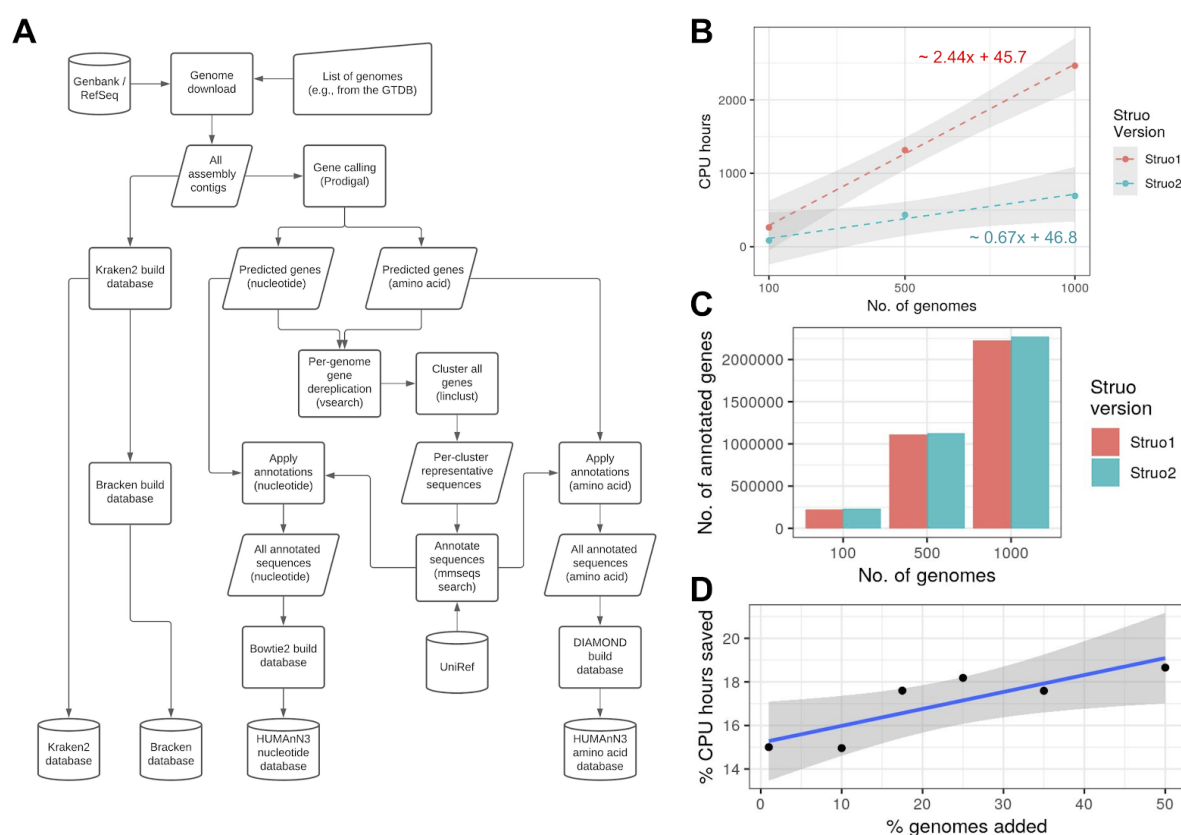
28 Metagenome profiling involves mapping reads to reference sequence databases and is
29 the standard approach for assessing microbial community taxonomic and functional composition
30 via metagenomic sequencing. Most metagenome profiling software includes “standard”
31 reference databases. For instance, the popular HUMAnN pipeline includes multiple databases
32 for assessing both taxonomy and function from read data (Franzosa *et al.*, 2018). Similarly,
33 Kraken2 includes a set of standard databases for taxonomic classification of specific clades
34 (e.g., fungi or plants) or all taxa (Wood *et al.*, 2019). While such standard reference databases
35 provide a crucial resource for metagenomic data analysis, they may not be optimal for the
36 needs of researchers. For example, a custom database that includes newly generated MAGs
37 can increase the percent of reads mapped to references (Youngblut *et al.*, 2020). The process
38 of making custom reference databases is often complicated and requires substantial
39 computational resources, which led us to create Struo for straight-forward custom metagenome
40 profiling database generation (de la Cuesta-Zuluaga *et al.*, 2020). However, Struo requires ~2.4
41 CPU hours per genome, which would necessitate >77,900 CPU hours (>9.1 years) if including
42 one genome per the 31,911 species in Release 95 of the Genome Taxonomy Database (GTDB)
43 (Parks *et al.*, 2018).

44 Struo2 generates Kraken2 and Bracken databases similarly to Struo (Lu *et al.*, 2017;
45 Wood *et al.*, 2019), but the algorithms diverge substantially for the time consuming step of gene
46 annotation required for HUMAnN database construction. Struo2 performs gene annotation by
47 clustering all gene sequences of all genomes using the *mmseqs2 linclust* algorithm, and then
48 each gene cluster representative is annotated via *mmseqs2 search* (Figure 1A; Supplemental
49 Methods) (Steinegger and Söding, 2017, 2018). In contrast, Struo annotates all non-redundant
50 genes of each genome with DIAMOND (Buchfink *et al.*, 2015). Struo2 utilizes snakemake and

51 conda, which allows for easy installation of all dependencies and simplified scaling to high
 52 performance computing systems (Köster and Rahmann, 2012).

53 Benchmarking on genome subsets from the GTDB showed that Struo2 requires ~0.67
 54 CPU hours per genome versus ~2.4 for Struo (Figure 1B). Notably, Struo2 annotates slightly
 55 more genes than Struo, possibly due to the sensitivity of the *mmseqs search* iterative search
 56 algorithm (Figure 1C). The use of *mmseqs2* allows for efficient database updating of new
 57 genomes and/or individual gene sequences via *mmseqs clusterupdate* (Figure S1); we show
 58 that this approach saves 15-19% of the CPU hours relative to generating a database from
 59 scratch (Figure 1D).

60 We used Struo2 to create publicly available Kraken2, Bracken, and HUMAnN3 custom
 61 databases from Release 95 of the GTDB (see Supplemental Methods). We will continue to
 62 publish these custom databases as new GTDB versions are released. The databases are
 63 available at <http://ftp.tue.mpg.de/ebio/projects/struo2/>. We also created a set of utility tools for
 64 generating NCBI taxdump files from the GTDB taxonomy and mapping between the NCBI and
 65 GTDB taxonomies. The taxdump files are utilized by Struo2, but these tools can be used more
 66 generally to integrate the GTDB taxonomy into existing pipelines designed for the NCBI
 67 taxonomy (available at https://github.com/nick-youngblut/gtdb_to_taxdump).



68 **Figure 1. Struo2 can build databases faster than Struo and can efficiently update the databases.** A) A
 69 general outline of the Struo2 database creation algorithm. Cylinders are input or output files, squares are
 70 processes, and right-tilted rhomboids are intermediate files. The largest change from Struo is the

71 utilization of mmseqs2 for clustering and annotation of genes. B) Benchmarking the amount of CPU hours
72 required for Struo and Struo2, depending on the number of input genomes. C) The number of genes
73 annotated with a UniRef90 identifier. D) The percent of CPU hours saved via the Struo2 database
74 updating algorithm versus *de novo* database generation. The original database was constructed from
75 1000 genomes. For B) and D), the grey regions represent 95% confidence intervals.

76 Data availability

77 Struo2 is available at <https://github.com/leylabmpi/Struo2>, the pre-built databases can be
78 found at <http://ftp.tue.mpg.de/ebio/projects/struo2/>, and utility tools are located at
79 https://github.com/nick-youngblut/gtdb_to_taxdump.

80 Acknowledgements

81 This study was supported by the Max Planck Society. We thank Albane Ruaud, Liam
82 Fitzstevens, Jacobo de la Cuesta-Zuluaga, and Jillian Waters for providing helpful comments on
83 an earlier version of this manuscript.

84 References

- 85 Buchfink, B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*,
86 **12**, 59–60.
- 87 de la Cuesta-Zuluaga, J. *et al.* (2020) Struo: a pipeline for building custom databases for
88 common metagenome profilers. *Bioinformatics*, **36**, 2314–2315.
- 89 Franzosa, E.A. *et al.* (2018) Species-level functional profiling of metagenomes and
90 metatranscriptomes. *Nat. Methods*, **15**, 962–968.
- 91 Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine.
92 *Bioinformatics*, **28**, 2520–2522.
- 93 Lu, J. *et al.* (2017) Bracken: estimating species abundance in metagenomics data. *PeerJ*
94 *Comput. Sci.*, **3**, e104.
- 95 Parks, D.H. *et al.* (2018) A standardized bacterial taxonomy based on genome phylogeny
96 substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.
- 97 Steinegger, M. and Söding, J. (2018) Clustering huge protein sequence sets in linear time. *Nat.*
98 *Commun.*, **9**, 2542.
- 99 Steinegger, M. and Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for
100 the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
- 101 Wood, D.E. *et al.* (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.
- 102 Youngblut, N.D. *et al.* (2020) Large-Scale Metagenome Assembly Reveals Novel
103 Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic
104 Diversity. *mSystems*, **5**.