Metaphor: facilitating the large-scale recovery of genomes from metagenomes

PRESENTER
Vinícius W. Salazar[1],
Babak Shaban[2], Maria Quiroga[2], Robert Turnbull[2], Edoardo Tescari[2], Vanessa Rossetto Marcelino[3], Heroen Verbruggen[4], Kim-Anh Lê Cao[1]
1 School of Mathematics & Statistics, University of Melbourne, 2 Melbourne Data Analytics Platform, 3 Hudson Institute of Biomedical Research, 4 School of Biosciences, University of Melbourne
@vinisalazar_ @vinisalazar

Metaphor is a general-purpose workflow for the assembly and binning of metagenomes. It advances existing metagenomic binning pipelines by the combination of two core features: the usage of multiple binning software along with a binning refinement step, and the possibility of defining groups for assembly and binning of samples. This latter feature allows users to pool samples together for the processes of assembly and binning (also called as coassembly or cobinning). We recommend doing this for samples which are biologically similar, *i.e.* that were sourced from the same community or environment, and as computational capacity allows. Metaphor is implemented with Snakemake [1], following best practices for workflow development, and the selection of software tools is informed by the Critical Assessment for Metagenome Interpretation [2]. To benchmark Metaphor, we ran it on simulated datasets with defined "ground-truth" reference genomes, allowing the comparison of different assembly and binning strategies. Metaphor is easily installable from Bioconda [3], compatible with various computing environments, and designed to be user-friendly, scalable, and automated.

General implementation: automate and scale your analysis with Snakemake.

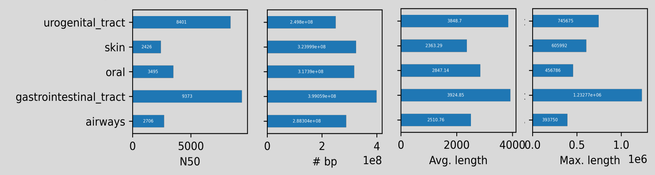
The workflow uses Snakemake, a scientific workflow management system, for calculation of a graph of jobs that can then be parallelized for faster execution. This can be done in multi-core machines, or combined with job schedulers such as SLURM and scaled to several computing nodes in high performance computing clusters. Metaphor is installed through the Bioconda software distribution channel, and Snakemake manages all dependencies automatically.

snakemake — A framework for reproducible data analysis
BIOCONDA
slurm workload manager

Quality Control: trim, filter and merge reads.

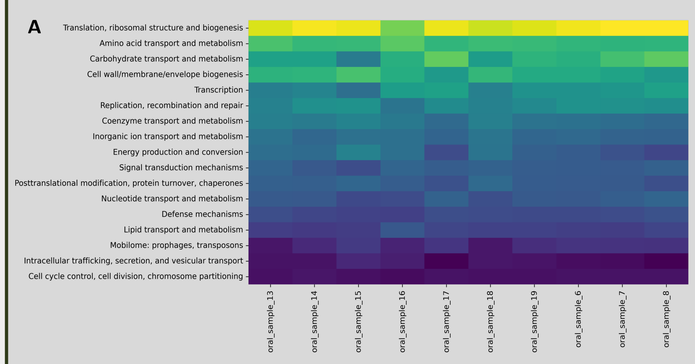
This module uses Cutadapt [4] to trim adapters and remove low-quality reads, and contaminants can be removed by mapping against a reference with MiniMap2 [5]. FastQC [6] reports are created for the raw and filtered reads, showing mean and per sequence scores (**A** and **B**, respectively), along with other metrics, and concatenated with MultiQC [7].

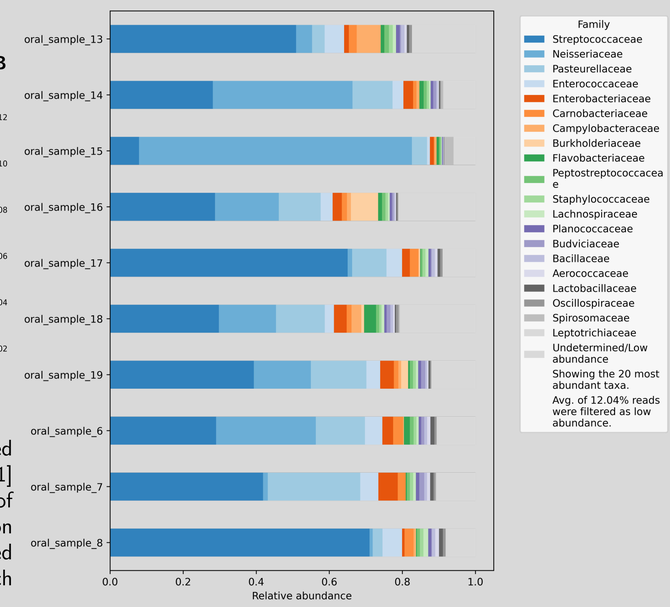
Assembly: assemble samples individually or in groups (coassembly).



Assembly is performed with MegaHit [8], and an assembly report summarises metrics for each sample or assembly group (panels above). Assemblies may also be evaluated with MetaQuast [9].

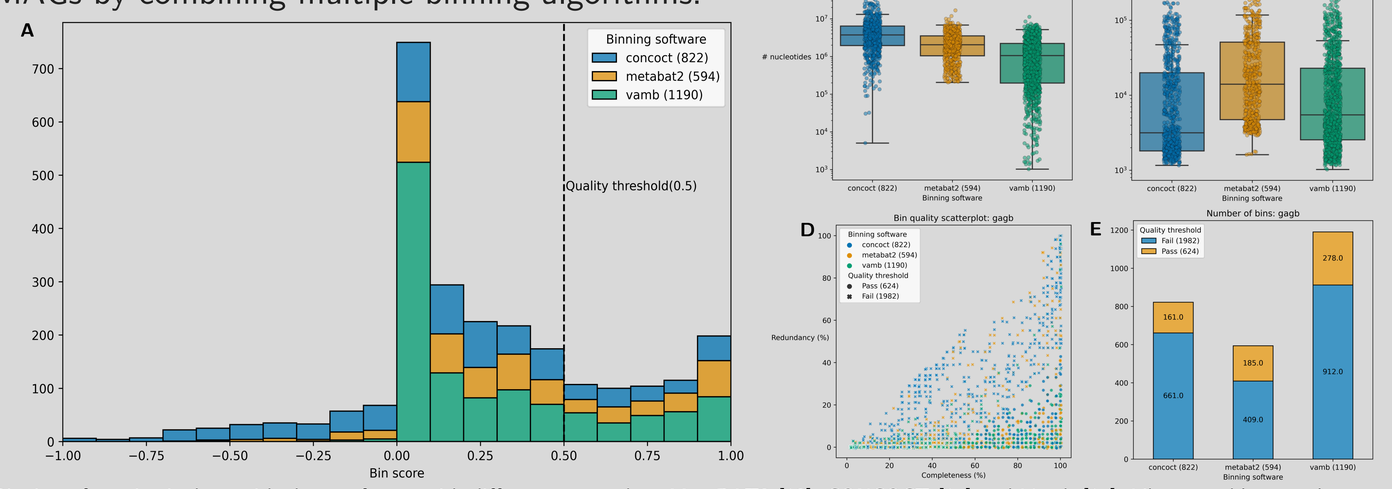
Annotation: obtain contig-level taxonomic and functional profiles.





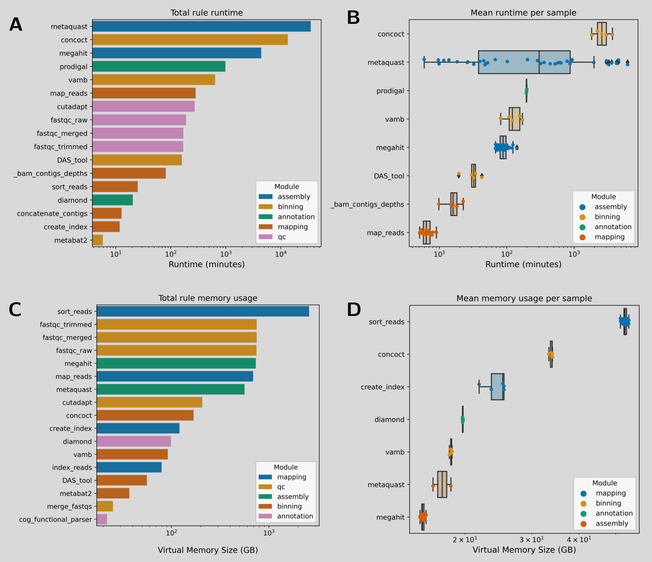
Metaphor uses Prodigal [10] to predict coding sequences from assembled contigs. The resulting amino acid sequences are aligned with Diamond [11] against the NCBI COG database [12]. Absolute and relative abundance of functional and taxonomic units is estimated using contig coverage information obtained from mapping with MiniMap2 [5]. Relative abundances are displayed as a heatmap of functional categories (**A**) and stacked bar plots for each taxonomic rank (**B** shows the Family rank).

Binning: generate high-quality and highly complete MAGs by combining multiple binning algorithms.

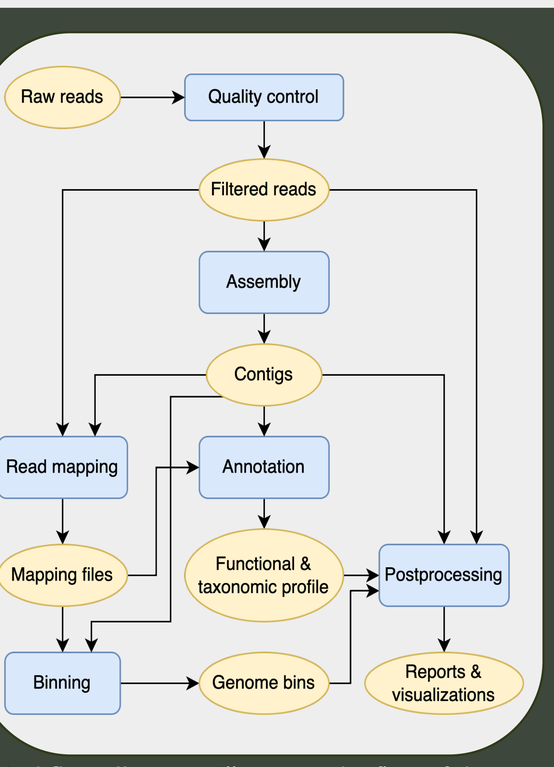


Binning of contigs is done with three software with different approaches: MetaBAT2 [13], CONCOCT [14] and Vamb [15]. Like assembly, samples may be binned individually or in groups, as the latter allows the binning software to leverage differential coverage information across samples. Bins are then refined by DAS Tool [16], which performs dereplication and quality check of the bins generated from each software. A score is then assigned to each bin based on completeness and redundancy (**A**), and bins below an arbitrary quality threshold are filtered. Additional reported metrics for each binning software are the size of each bin (**B**), the contiguity (as N50) (**C**), the completeness and redundancy (**D**), and the number of bins passing the quality threshold (**E**). In the plot titles, 'gagb' refers to the assembly and binning strategy: group assembly and group binning (of the CAMI human gut dataset). The final set of bins may then be annotated with Prokka [17].

Postprocessing: identify computational bottlenecks in the pipeline execution.



Upon completion of execution, Metaphor generates a graphical report of the total and per-sample runtime (**A** and **B**) and memory usage (**C** and **D**).

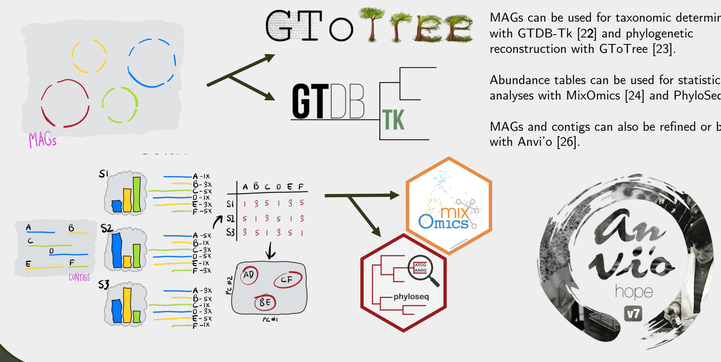


Workflow diagram: illustrates the flow of data (yellow oval shapes) between the modules (blue square shapes) of Metaphor.

Steps	Metaphor	ATLAS [18]	MetaWRAP [19]	nf-core/MAG [20]	MAGNETO [21]
Preprocessing					
Reads trimming	✓	✓	✓	✓	✓
Contamination	✓	✓	✓	✓	✓
Assembly					
Coassembly possible	✓		✓	✓	✓
Compute sets to coassemble	✓			✓	
Binning					
Cobinning possible	✓		✓	✓	✓
Multiple binning software	✓	✓	✓		
Bin refinement	✓	✓	✓		
Bin reassembly		✓	✓		
Postprocessing					
MAGs quality check	✓	✓	✓	✓	✓
Dereplication step	✓	✓	✓	✓	✓
Genome annotation	✓	✓	✓	✓	✓
Gene catalogue	✓			✓	✓
Reproducibility					
Workflow management	✓	✓		✓	✓
Packages management	✓	✓		✓	✓

Workflow features. Metaphor when compared with four other modern metagenomics pipelines. Table adapted from [21].

The output of Metaphor may be directly used in a range of downstream applications.



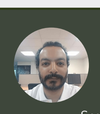
MAGs can be used for taxonomic determination with GTDB-Tk [22] and phylogenetic reconstruction with GToTree [23].

Abundance tables can be used for statistical analyses with MixOmics [24] and PhyloSeq [25].

MAGs and contigs can also be refined or binned with Anvi'o [26].

References

1. Mölder, F. et al. Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).
2. Meyer, F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. Nat Methods 19, 429–440 (2022).
3. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15, 475–476 (2018).
4. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
5. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
6. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. (2020).
7. Ewels, P., et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016) doi:10.1093/bioinformatics/btw354.
8. Li, D., et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
9. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088–1090 (2016).
10. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11, 119 (2010).
11. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nature Methods (2014) doi:10.1038/nmeth.3176.
12. Galperin, M. Y. et al. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49, D274–D281 (2021).
13. Kang, D. D. et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019, 1–13 (2019).
14. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nature Methods 11, 1144–1146 (2014).
15. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. Nature Biotechnology (2021) doi:10.1038/s41587-020-00777-4.
16. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3, 836–843 (2018).
17. Seemann, T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30, (2014).
18. Kieser, S., et al. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinformatics 21, 1–8 (2020).
19. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline fo r genome-resolved metagenomic data analysis. Microbiome 6, 158 (2018).
20. Krakau, S., et al. nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. (2021) doi:10.1101/2021.08.29.458094.
21. Churcheward, B., Millet, M., Bilodeau, A., Fertin, G. & Chaffron, S. MAGNETO: An Automated Workflow for Genome-Resolved Metagenomics. mSystems 0, e00432-22 (2022).
22. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927 (2019).
23. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics (2019) doi:10.1093/bioinformatics/btz188.
24. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology (2017)
25. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8, e61217 (2013).
26. Community-led, integrated, reproducible multi-omics with anvi'o | Nature Microbiology. https://www.nature.com/articles/s41564-020-00834-3.



In loving memory of Bobbie Shaban, an esteemed colleague and friend.

Scan for poster and documentation.

This work was funded by a Melbourne Research Scholarship.

THE UNIVERSITY OF MELBOURNE