Analysis of Complex Metagenomes with MetScale™ Workflows

#78

M. Scholz • N. Keplinger • C. Grahlmann • C. Hulme-Lowe • M. Isbell • T. Treangen • K. Ternus

ABSTRACT

Metagenomics is a powerful tool that allows researchers to gain insights into the taxonomic and functional content of complex microbial communities without the need for culturing. As metagenomics has expanded, there has been a corresponding increase in available taxonomic classification tools and reference databases to evaluate such sequences. However, positive calls at the strain, species, or genus levels may vary significantly among taxonomic classification tools. Furthermore, inconsistencies among the reference databases used by these tools can lead to false positives and/or false negatives, resulting in database-linked biases.

To characterize and address these gaps, we developed MetScale™ workflows to allow users to execute multiple open source metagenomic tools and reference databases at one time and to assist users in differentiating true positive from false positive signals. MetScale™ can run online or offline on an air-gapped system, and its workflows include tools for Illumina® read filtering, assembly, taxonomic classification, and functional inference.

The MetScale™ final report summarizes the results for each sample analyzed and describes the species identified by different taxonomic classification tools and reference databases.

https://github.com/signaturescience/metscale

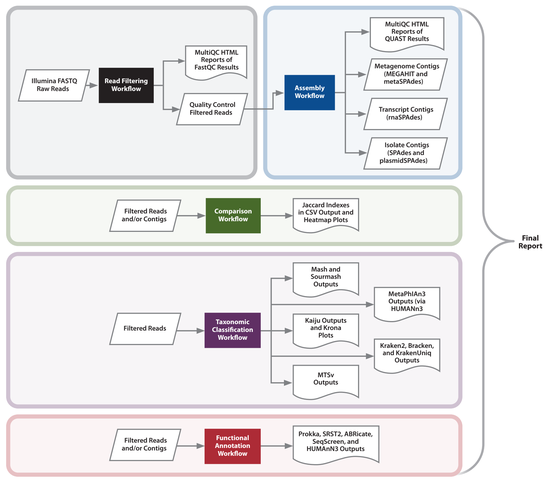

Figure 1: Flow chart of MetScale™ workflows. Results for each sample are combined into a single final report.

DATABASE QUERY TOOL (DQT)

The accuracy and sensitivity of a taxonomic classifier is limited by the comprehensiveness of its reference database, but it is not always easy to see what organisms are present within different reference databases. The Database Query Tool (DQT) enables manual queries to assess the taxonomic composition of databases used by metagenomic classification tools. This allows users to explore if the absence of an identified species was likely a true negative, or if it was likely a false negative due to that species being absent from the reference database used with the taxonomic classification tool. MetScale™ implements the DQT with a compiled database generated from the default references for each of the classification tools. The database can be re-compiled with new or custom reference databases, if desired.

ANALYSIS OF A CONSTRUCTED METAGENOME

The Shakya synthetic metagenome[18] is a constructed meta–genome consisting of 64 fully sequenced bacteria, archeaea and fungi. Sequence data from this synthetic community has been used to assess the accuracy of taxonomic classification tools (e.g. McIntyre et al. 2017)[19]. MetScale™ incorporates the results of assembly, taxonomic, and functional classification tools into a single report. When compared to the known truth data, the value of combining multiple pipelines can be demonstrated.

Available Workflows:

Because MetScale is a workflow management system, it is able to incorporate multiple tools, generally divided into 5 categories:

1 Quality Control

MetScale incorporates read trimming, read QC and Kmer counting with Trimmomatic[1], FastQC[2], and Khmer[3], respectively. Quality control data are consolidated into a single report using MultiQC[4]. Multiple parameters can be configured for each tool (trimming quality, minimum length, etc.)

2 Assembly

There are two main assembly tools implemented in MetScale: SPAdes[5] and MEGAHIT[6]. Assembly quality can be checked with QUAST[7]/metaQUAST[8], and reports are again incorporated into MultiQC reports, as well as individual reports.

3 Comparison

Assemblies can be compared with calculated Jaccard Indexes across all configured trimming and assembly parameters, to determine the level of difference between assembly methodologies. Comparisons are performed using Sourmash[9].

4 Taxonomic Classification

Taxonomic classification is available in MetScale from 6 separate tools, Mash[9], Kaiju[10], Kraken2[11], KrakenUniq[12], and MTSv[13]. MetScale can process reads and/or assembled contigs for each of these steps. For all tools, MetScale is configured to download selected database for analysis. (See DQT section)

5 Functional Classification

There are separate tools available within Functional Classification for Contigs and Read based classification. Functional genes can be predicted for contigs using either Prokka[14] or Abricate[15]. Prokka functions by predicting open reading frames (ORF) and predicting function using a database search/ voting strategy for each ORF. Abricate screens for antibiotic resistance genes. Read based functional classification can be performed using either Humann3[16] or SRST2[17].

Data Gathering:

MetScale creates a final report by tying together all analyses that have been performed to generate a single HTML-based page allowing the user to navigate all results, as well as combining all taxonomic tool outputs into a single table to identify genera or species most commonly identified across analyses.

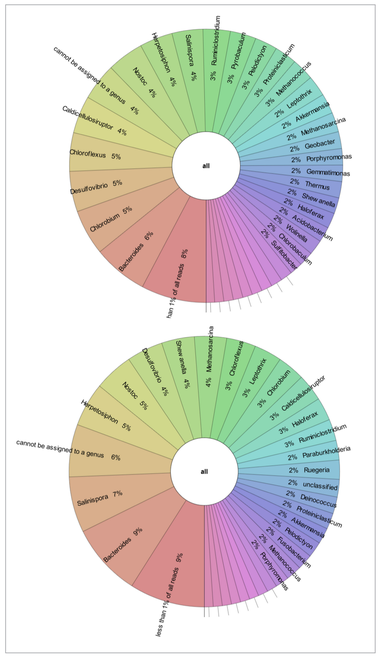

Figure 2: Example output for Shakya dataset analyzed with Metscale™. The Krona plots above show the results of Kaiju classification of the assembled contigs of the same dataset, using either MetaSPAdes or MEGAHIT at the genus level. While the same species are identified in both analyses, the relative abundance varies, illustrating the value in utilizing multiple tools for each step in a pipeline.

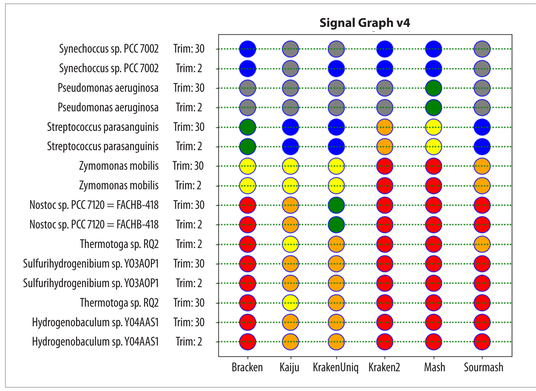

Figure 3: Signal graph generated from the MetScale™ analysis of the Shakya dataset. The top 8 species from each unique combination of parameters (in this case trimming at Q=2 or Q=30) are identified. This plot summarizes the outputs of all taxonomic classifiers run. Color of the circles indicates the level of supporting evidence for that species.

	Species Signal	KrakenUniq kmers	Kraken2 reads	Bracken reads	Kaiju reads	Sourmash f match	Mash identity
Red	Very Strong	>10,000	>100,000	>100,000	>100,000	>0.60	>0.95
Orange	Strong	5,000-10,000	30,000-100,000	30,000-100,000	30,000-100,000	0.20-0.60	0.90-0.95
Yellow	Moderately Strong	2,000-5,000	10,000-30,000	10,000-30,000	10,000-30,000	0.15-0.20	0.85-0.90
Green	Moderate	1,000-2,000	1,000-10,000	1,000-10,000	1,000-10,000	0.10-0.15	0.80-0.85
Blue	Weak	500-1,000	100-1,000	100-1,000	100-1,000	0.05-0.10	0.75-0.80
Grey	Very Weak	0-500	0-100	0-100	0-100	0-0.05	0-0.75
White	No Species Signal	0	0	0	0	0	0

Table 1: Rubric for classification of strength of a species signal from each of the classification tools implemented in MetScale™.

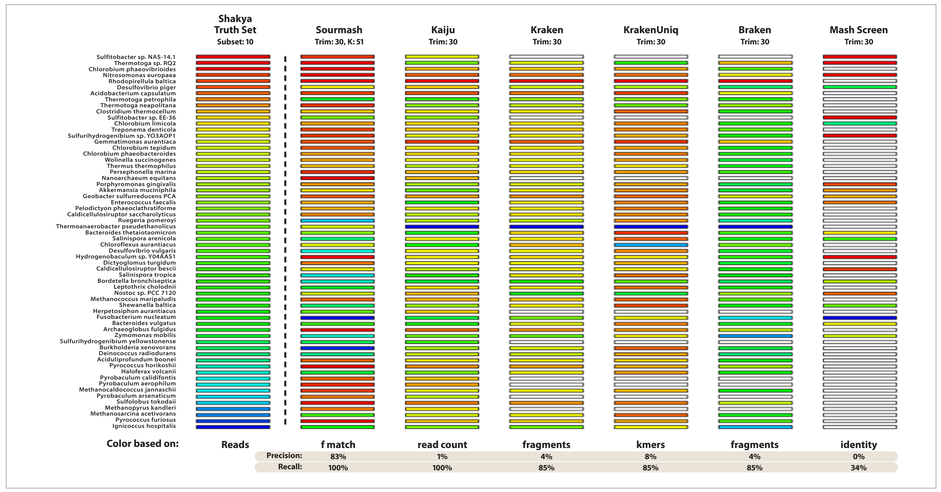

Figure 4: Heatmap display of all bacterial members of Shakya dataset, including truth data. Shading is the same as Figure 3. White boxes indicate no support from that tool for the presence of species of interest. Use of the Database Query Tool (DQT) allows us to determine that the absence is due to the species not being present in the reference database used for those tools.

REFERENCES

1. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," Bioinformatics, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
2. "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Oct. 02, 2022).
3. M. R. Crusoe et al., "The khmer software package: enabling efficient nucleotide sequence analysis," F1000Res, vol. 4, p. 900, Sep. 2015, doi: 10.12688/f1000research.6924.1.
4. P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," Bioinformatics, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
5. A. Bankevich et al., "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," Journal of Computational Biology, vol. 19, no. 5, pp. 455–477, 2012, doi: 10.1089/cmb.2012.0021.
6. D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," Bioinformatics, vol. 31, no. 10, pp. 1674–1676, 2015, doi: 10.1093/bioinformatics/btv033.
7. A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: quality assessment tool for genome assemblies," Bioinformatics, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, doi: 10.1093/bioinformatics/btt086.
8. A. Mikheenko, V. Saveliev, and A. Gurevich, "MetaQUAST: evaluation of metagenome assemblies," Bioinformatics, vol. 32, no. 7, pp. 1088–1090, Apr. 2016, doi: 10.1093/bioinformatics/btv697.
9. N. T. Pierce, L. Irber, T. Reiter, P. Brooks, and C. T. Brown, "Large-scale sequence comparisons with sourmash," F1000Res, vol. 8, p. 1006, Jul. 2019, doi: 10.12688/f1000research.19675.1.
10. P. Menzel, K. L. Ng, and A. Krogh, "Fast and sensitive taxonomic classification for metagenomics with Kaiju," Nat Commun, vol. 7, no. 1, Art. no. 1, Apr. 2016, doi: 10.1038/ncomms11257.
11. D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," Genome Biology, vol. 20, no. 1, p. 257, Nov. 2019, doi: 10.1186/s13059-019-1891-0.
12. F. P. Breitwieser, D. N. Baker, and S. L. Salzberg, "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts," Genome Biology, vol. 19, no. 1, p. 198, Nov. 2018, doi: 10.1186/s13059-018-1568-0.
13. J. M. Wood, N. K. Singh, L. Guan, A. Seuylemezian, J. N. Benardini, and K. Venkateswaran, "Performance of Multiple Metagenomics Pipelines in Understanding Microbial Diversity of a Low-Biomass Spacecraft Assembly Facility," Front Microbiol, vol. 12, p. 685254, Sep. 2021, doi: 10.3389/fmicb.2021.685254.
14. T. Seemann, "Prokka: rapid prokaryotic genome annotation," Bioinformatics, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.
15. T. Seemann, "ABRicate." Oct. 03, 2022. Accessed: Oct. 03, 2022. [Online]. Available: https://github.com/tseemann/abricate
16. "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3 | eLife." https://elifesciences.org/articles/65088 (accessed Oct. 03, 2022).
17. "SRST2: Rapid genomic surveillance for public health and hospital microbiology labs | Genome Medicine | Full Text." https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-014-0090-6 (accessed Oct. 03, 2022).
18. M. Shakya, C. Quince, J. H. Campbell, Z. K. Yang, C. W. Schadt, and M. Podar, "Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities," Environ Microbiol, vol. 15, no. 6, pp. 1882–1899, Jun. 2013, doi: 10.1111/1462-2920.12086.
19. A. B. R. McIntyre et al., "Comprehensive benchmarking and ensemble approaches for metagenomic classifiers," Genome Biology, vol. 18, no. 1, p. 182, Sep. 2017, doi: 10.1186/s13059-017-1299-7.

ACKNOWLEDGEMENTS

This work was built off development and contributions from the Lab for Data Intensive Biology at the University of California, Davis, the Treangen lab at Rice University, the Budowle lab at the University of North Texas Health Science Center, and the University of Virginia's Bioinformatics Core. We would also like to thank all of the developers of the open source code, tools, and databases that are incorporated into the MetScale pipelines (https://github.com/signaturescience/metscale/blob/master/DEPENDENCY_LICENSES).

Signature Science, LLC
8329 North Mopac Expressway
Austin, TX 78759
www.signaturescience.com

METSCALE™
Snakemake Workflows to Scale Metagenomic Analyses

signature science® LLC