The Journal of Open Source Software

YACHT: Software for an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample

Maksym Lupei [1*], **Shaopeng Liu** [2*], **Chunyu Ma** [1*], **Adam Park** [1*], **Omar Hesham Rady** [1*], **Mahmudur Rahman Hera** [1*], **Judith S. Rodriguez** [2*], **Stephanie J. Won** [3*], **and David Koslicki** [1,2,3¶]

1 School of Electrical Engineering and Computer Science, Pennsylvania State University, United States of America **2** Huck Institutes of the Life Sciences, Pennsylvania State University, United States of America **3** Department of Biology, Pennsylvania State University, United States of America ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.09066

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @aboffin
- @Vini2
- @anuradhawick

Submitted: 22 May 2025
Published: 02 May 2026

Summary

Identifying genomes in metagenomic samples can be complicated by taxonomic profiling tools that lack uncertainty quantification and rely on incomplete reference databases. YACHT (**Y**es/No **A**nswers to **C**ommunity membership via **H**ypothesis **T**esting) introduces a k-mer sketching based statistical framework that incorporates average nucleotide identity (ANI) and coverage, the portion of k-mers observed for a microbe's genome detected in a sample, to detect genetic similarity between reference and sample genomes using binomial hypothesis testing on exclusive k-mers to confidently determine genome presence/absence (Koslicki et al., 2024). This paper describes the software implementation of this methodology as a command-line tool that detects low-abundant species while controlling the false-negative rate, making it applicable to functional profiling, metatranscriptomics, and clinical microbiome analysis despite incomplete genomes and variable coverage. YACHT is developed with C++ and Python and depends on `sourmash` (Irber et al., 2024) for k-mer extraction and management.

Statement of need

Accurately identifying low-abundance microbial communities remains a significant challenge in metagenomics. Current methods rely on arbitrary filter thresholds that, even when applied, produce results skewed by sequencing errors and evolutionary processes, compromising profiling accuracy and leading to misinterpretations (Jia et al., 2022; Schloss, 2020). The lack of a systematic credibility framework can undermine researcher confidence, a problem compounded by incomplete reference databases and variable coverage.

Metagenomic methods depend on reference databases that are often incomplete and misaligned with taxonomic frameworks, leaving evolutionarily diverged microbes undetected and causing profiling inaccuracies (Kunin et al., 2008; Loeffler et al., 2020; Vanessa R. Marcelino et al., 2020; Schlaberg et al., 2017). Addressing this requires analytical frameworks that incorporate genome similarity metrics, though coverage presents an additional challenge to reliable microbial detection.

Coverage is crucial for detecting low-abundance microbes, which are often misinterpreted as noise due to limited sequencing depth (Mande et al., 2012; Meyer et al., 2022; Sczyrba et al., 2017; Shakya et al., 2013). The lack of guidelines for biologically meaningful coverage thresholds introduces subjectivity, making dynamic coverage thresholds essential. Yet even

with adequate coverage and reliable genome references, controlling statistical errors remains a major challenge.

Existing metagenomic methods lack the statistical rigor to control false positives and false negatives effectively, where high false positive rates misrepresent microbial composition and false negative rates cause researchers to overlook biologically important taxa (Jousset et al., 2017). Incomplete reference databases, sequencing errors, and evolutionary divergence between reference and sample genomes further complicate statistical error rates, making a multifaceted statistical approach essential to capture microbial profiling accurately.

YACHT addresses these challenges through hypothesis testing that accounts for evolutionary sequence divergence and incomplete sequencing depth utilizing genome similarity and minimum sequencing depth parameters. It employs the FracMinHash sketching technique (Irber Jr, 2020; Irber et al., 2022), an alignment-free k-mer approach, facilitating fast and accurate detection of low abundance taxa with a user-defined false negative rate. YACHT is applicable to functional profiling, metatranscriptomic studies (Vanesa R. Marcelino et al., 2019), metabolic potential analyses (Pereira-Marques et al., 2024; Ward et al., 2018), and the characterization of low abundant clinical metagenomic samples such as skin (Godlewska et al., 2020), reducing reliance on arbitrary thresholds and distinguishing genuine artifacts from "noise" with statistical confidence.

Workflow

The YACHT workflow involves four primary steps. First, `yacht sketch` samples compact representations of reference genomes. Second, `yacht train` preprocesses the reference genomes, merging those with high ANI into a single representative. Third, `yacht run` executes the core YACHT algorithm to perform hypothesis testing and determine the membership of organisms. Finally, `yacht convert` transforms the results into popular output formats like CAMI, BIOM, and GraphPhlAn.

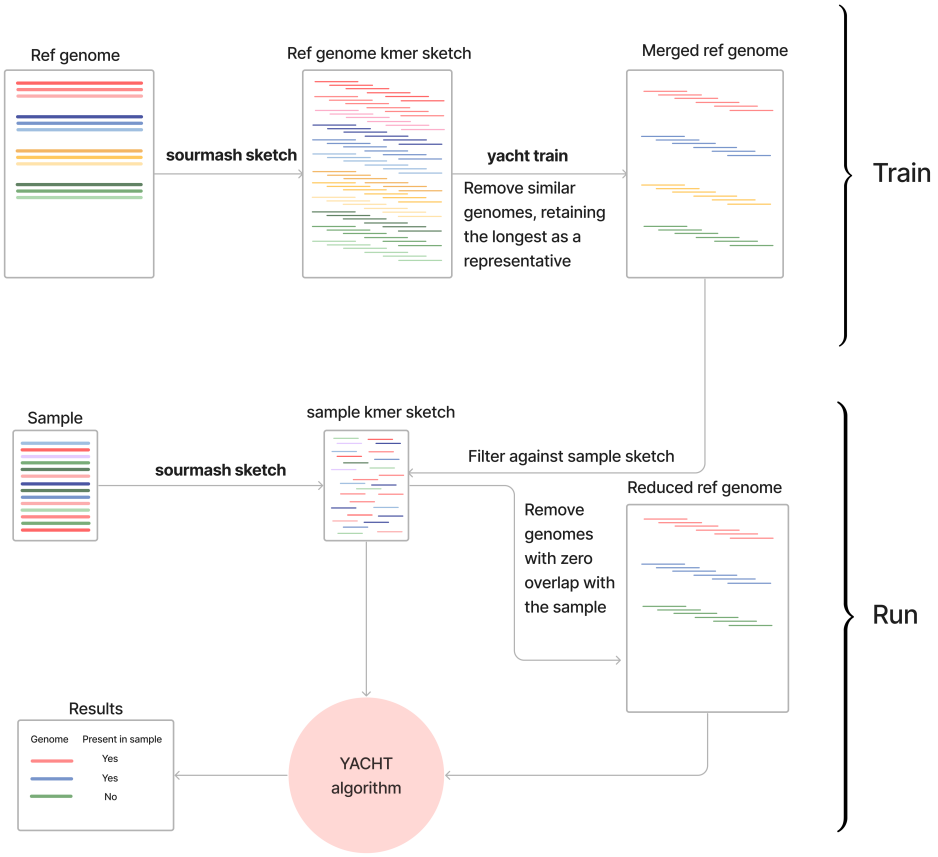

Figure 1: The YACHT workflow illustrated with the four primary stages: sketching, training, running, and converting.

As outlined in the workflow in **Figure 1**, YACHT requires two primary inputs: a pre-trained reference configuration (JSON) and a sketched sample signature. See the repository for a detailed step-by-step workflow.

Output examples

The `yacht run` output provides probabilistic decisions on organism presence or absence, as shown in **Table 1** below. For each organism, columns like `num_matches` and `acceptance_threshold` are reported, indicating the number of k-mers found and the minimum required to be considered present, respectively. The `Presence` column then reports `TRUE` or `FALSE` based on this comparison.

Use case examples

We present the three use case examples to demonstrate the application of YACHT for identifying taxonomy in microbiome studies: (i) analyzing low-abundance metagenomic samples that

Organism	Presence	num_matches	acceptance_threshold	alt_confidence_mut_rate
Sediminispirochaeta	TRUE	2572	895	0.053008659
Natronobacterium	TRUE	700	638	0.053534755
Echinicola	FALSE	244	978	0.052885411

Table 1: YACHT results for Sediminispirochaeta, Natronobacterium, and Echinicola showing a subset of output columns: whether the organism passed the presence threshold (Presence), the number of exclusive k-mer matches (num_matches), the expected minimum number of matches (acceptance_threshold), and an alternative confidence estimate for the mutation rate (alt_confidence_mut_rate). Note that Echinicola is not reported as present, while Sediminispirochaeta and Natronobacterium are present meeting the acceptance threshold. Results were generated using the MBARC-26 dataset (SRA: SRR6394747 by @Singer2016MockCommunity) with YACHT parameters: k-size of 31, minimum coverage of 0.05, and ANI threshold of 0.95.

are common in clinical settings, (ii) performing MAG fishing to detect specific metagenomic-assembled genomes, and (iii) evaluating synthetic microbial communities to identify the presence of specific organisms.

Low abundance samples: YACHT can analyze metagenomic samples with low microbial DNA concentrations common in clinical and environmental studies. Using a human skin metagenomics sample, we show that ANI threshold and k-size markedly influence species specificity. See Low abundance samples.

Metagenomic-assembled genome (MAG) fishing: Using a single MAG as a training reference database, YACHT searches for specific MAGs within a sample. Applied to two skin metagenomic samples, result shows that detection is sensitive to sequencing depth, coverage, and parameter choice. See MAG fishing.

Synthetic metagenomes: YACHT verifies the presence of designed microbes in mock microbial communities. Higher ANI thresholds recover expected genomes while lower thresholds introduce false positives, demonstrating how ANI and minimum coverage parameters affect sensitivity and specificity. See Synthetic metagenomes

Acknowledgements

We thank the contributors and collaborators who supported the development of YACHT. This work was supported in part by the National Institutes of Health (NIH) under grant number 5R01GM146462-03.

References

Godlewska, U., Brzoza, P., Kwiecień, K., Kwitniewski, M., & Cichy, J. (2020). Metagenomic studies in inflammatory skin diseases. *Current Microbiology*, *77*, 3201–3212. https://doi.org/10.1007/s00284-020-02163-4

Irber Jr, L. C. (2020). *Decentralizing indices for genomic data*. University of California, Davis.

Irber, L., Brooks, P. T., Reiter, T. E., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., & Brown, C. T. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *bioRxiv*. https://doi.org/10.1101/2022.01.11.475838

Irber, L., Pierce-Ward, N. T., Abuelanin, M., Alexander, H., Anant, A., Barve, K., Baumler, C., Botvinnik, O., Brooks, P., Dsouza, D., & others. (2024). Sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *Journal of Open Source Software*, *9*(98), 6830. https://doi.org/10.21105/joss.06830

Jia, Y., Zhao, S., Guo, W., Peng, L., Zhao, F., Wang, L., Fan, G., Zhu, Y., Xu, D., Liu, G., & others. (2022). Sequencing introduced false positive rare taxa lead to biased microbial community diversity, assembly, and interaction interpretation in amplicon studies. *Environmental Microbiome*, *17*(1), 43. https://doi.org/10.1186/s40793-022-00436-y

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M. C., Rivett, D. W., Salles, J. F., & others. (2017). Where less may be more: How the rare biosphere pulls ecosystems strings. *The ISME Journal*, *11*(4), 853–862. https://doi.org/10.1038/ismej.2016.174

Koslicki, D., White, S., Ma, C., & Novikov, A. (2024). YACHT: An ANI-based statistical test to detect microbial presence/absence in a metagenomic sample. *Bioinformatics*, *40*(2), btae047. https://doi.org/10.1093/bioinformatics/btae047

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, *72*(4), 557–578. https://doi.org/10.1128/MMBR.00009-08

Loeffler, C., Karlsberg, A., Martin, L. S., Eskin, E., Koslicki, D., & Mangul, S. (2020). Improving the usability and comprehensiveness of microbial databases. *BMC Biology*, *18*, 1–6. https://doi.org/10.1186/s12915-020-0756-z

Mande, S. S., Mohammed, M. H., & Ghosh, T. S. (2012). Classification of metagenomic sequences: Methods and challenges. *Briefings in Bioinformatics*, *13*(6), 669–681. https://doi.org/10.1093/bib/bbs054

Marcelino, Vanessa R., Clausen, P. T., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*, 1–15. https://doi.org/10.1186/s13059-020-02014-2

Marcelino, Vanesa R., Irinyi, L., Eden, J.-S., Meyer, W., Holmes, E. C., & Sorrell, T. C. (2019). Metatranscriptomics as a tool to identify fungal species and subspecies in mixed communities–a proof of concept under laboratory conditions. *IMA Fungus*, *10*, 1–10. https://doi.org/10.1186/s43008-019-0012-8

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., & others. (2022). Critical assessment of metagenome interpretation: The second round of challenges. *Nature Methods*, *19*(4), 429–440. https://doi.org/10.1038/s41592-022-01431-4

Pereira-Marques, J., Ferreira, R. M., & Figueiredo, C. (2024). A metatranscriptomics strategy for efficient characterization of the microbiome in human tissues with low microbial biomass. *Gut Microbes*, *16*(1), 2323235. https://doi.org/10.1080/19490976.2024.2323235

Schlaberg, R., Chiu, C. Y., Miller, S., Procop, G. W., Weinstock, G., Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology, & Microbiology Resource Committee of the College of American Pathologists. (2017). Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Archives of Pathology and Laboratory Medicine*, *141*(6), 776–786. https://doi.org/10.5858/arpa.2016-0539-RA

Schloss, P. D. (2020). Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data. *bioRxiv*. https://doi.org/10.1101/2020.12.11.422279

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., & others. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., & Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, *15*(6), 1882–1899. https://doi.org/10.1111/1462-2920.12086

Ward, L. M., Shih, P. M., & Fischer, W. W. (2018). MetaPOAP: Presence or absence of metabolic pathways in metagenome-assembled genomes. *Bioinformatics*, *34*(24), 4284–4286. https://doi.org/10.1093/bioinformatics/bty510