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, USA **2** Huck Institutes of the Life Sciences, Pennsylvania State University, USA **3** Department of Biology, Pennsylvania State University, USA ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @aboffin
- @Vini2
- @anuradhawick

Submitted: 22 May 2025
Published: unpublished

Summary

In metagenomics, identifying genomes present in a sample is an important initial task, but is complicated by taxonomic profiling tools lacking uncertainty quantification and using incomplete reference databases missing exact genome matches. YACHT (**Y**es/No **A**nswers to **C**ommunity membership via **H**ypothesis **T**esting) (Koslicki et al., 2024) is a command-line tool for taxonomic profiling that uses binomial hypothesis testing on exclusive k-mers to confidently determine genome presence/absence in a metagenomic sample. YACHT assists in discovering rare microbiomes by identifying low-abundant species missed in other taxonomic profiling approaches while also controlling the false negative rate. Its statistical model overcomes challenges in sequencing coverage and incomplete genomes, making it ideal for diverse metagenomic applications, including functional profiling, metatranscriptomics, and clinical microbiome analysis.

YACHT presents a robust, k-mer sketching-based statistical framework for accurately detecting genetic similarity between the reference database and the metagenomic sample by incorporating evolutionary sequence divergence through the average nucleotide identity (ANI) and sequencing coverage to enable efficient detection of sampled genomes. The workflow for YACHT includes the following commands. To begin, `yacht sketch` creates reduced representation "sketches" of the reference and sample datasets enabling swift comparisons. Then, `yacht train` is used to find a representative of closely related reference genomes using ANI. Lastly, `yacht run` uses the YACHT algorithm to perform hypothesis testing and identify the presence or absence of species. YACHT is developed with C++ and Python and depends on `sourmash` (L. Irber et al., 2024), a program for extracting and managing k-mers.

Statement of need

Accurately identifying and characterizing microbial communities with low relative abundance is a significant challenge in metagenomics. The current profiling-based practice involves setting arbitrary filter thresholds or discarding low-abundance data without robust justification, which can compromise profiling accuracy and lead to misinterpretations (Jia et al., 2022; Schloss, 2020). Even with such filtering, the results remain inherently arbitrary because they are influenced by biological complexities such as sequencing errors and evolutionary processes. The lack of a systematic approach to establishing credibility in these results diminishes researchers' confidence in biologically informed methods for identifying rare microorganisms,

thereby undermining metagenomic studies. Moreover, these difficulties are exacerbated by the incompleteness of reference databases and the variability in sequencing coverage depth, underscoring the need for statistically credible approaches.

Metagenomic methods rely on existing genome references to detect and classify microbial organisms. However, these reference databases are often incomplete, and conventional metrics may not always align with traditional taxonomic frameworks that account for genomic changes. Consequently, microbes that carry mutations or have diverged evolutionarily can remain undetected, causing inaccuracies in microbial community profiling and misinterpretation of data (Kunin et al., 2008; Loeffler et al., 2020; Vanessa R. Marcelino et al., 2020; Schlaberg et al., 2017). Hence, analytical frameworks need to incorporate genome similarity metrics to capture the full breadth of microbial diversity and to provide accurate, interpretable microbiome dynamics. However, incomplete databases alone do not account for all metagenomic challenges; sequence coverage depth also contributes to the resolution and reliability of microbial detection and characterization.

Sequence coverage depth, defined as the portion of a microbe's genome detected in a sample, is crucial for detecting low-abundance microbes. However, sequencing processes often fail to achieve complete coverage of all genomes in a sample due to limited sequencing depth. As a result, rare or low-abundance taxa may exhibit low sequence coverage, leading to their misinterpretation as noise rather than genuine observations (Mande et al., 2012; Meyer et al., 2022; Sczyrba et al., 2017; Shakya et al., 2013). Furthermore, the lack of guidelines for establishing a biologically meaningful coverage depth threshold introduces subjectivity and inconsistency in the metagenomic analyses. Therefore, implementing dynamic coverage depth thresholds tailored to varying abundance levels is essential for delivering accurate metagenomic studies. Yet, even if we address coverage depth and incomplete genome reference problems, ensuring proper control over statistical errors remains another major challenge.

Existing metagenomic methods lack the statistical rigor to control false positives and false negatives effectively. High false positive rates misrepresent microbial composition and lead to biased conclusions, undermining research reliability. Conversely, false negative rates cause researchers to overlook important taxa, especially those in low abundance that often carry significant biological importance (Jousset et al., 2017). Incomplete reference databases, sequencing errors, and evolutionary divergence between reference and sample genomes further complicate these challenges. Therefore, maintaining appropriate control over these statistical error rates is critical to ensure more confident, reliable biological inferences and minimize the risk of misinterpretation. While limitations in reference database, sequence coverage depth and balance of statistical error pose significant challenges, the complexity of metagenomic analysis demands a multifaceted approach to capture microbial profiling accurately.

To address these challenges, YACHT offers a statistical framework that can accurately determine the presence or absence of microbial genome in a sample through hypothesis testing. The algorithm's mathematical model accounts for evolutionary sequence divergence and incomplete sequencing depth by utilizing genome similarity and minimum sequencing depth parameters. It employs the FracMinHash sketching technique (Irber Jr, 2020; L. C. Irber et al., 2022), an alignment-free k-mer approach, facilitating fast and accurate genome detection that can efficiently process large datasets. YACHT ensures precise detection of low abundance taxa with a user-defined false negative rate, minimizing the risk of misinterpretation of the result. Our approach can be used for other metagenomic applications such as 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). YACHT enhances metagenomic analysis by offering reduced reliance on arbitrary thresholds, improving the interpretability of the result without compromising biological relevance, and allowing researchers to differentiate between genuine artifacts from "noise" with statistical confidence.

Workflow

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

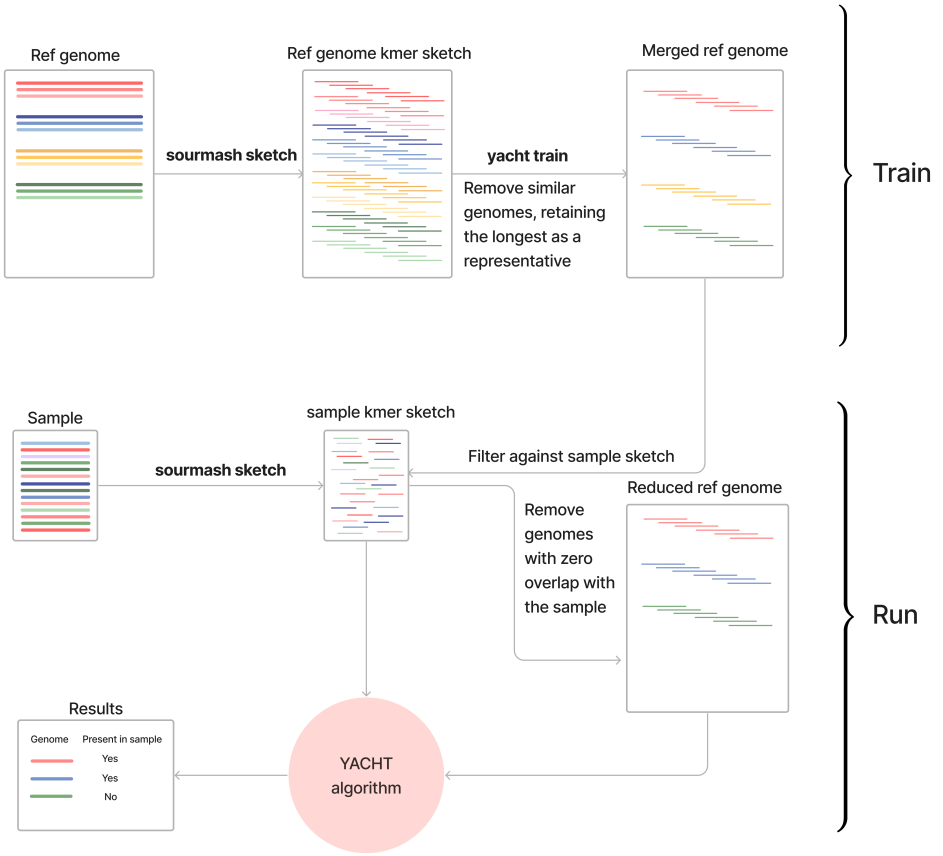

Figure 1: The YACHT workflow illustrated with the four primary stages: sketching, training, running, and converting.

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.

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 are reported. For each species, the following are shown as a subset of the output: 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) are shown. 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. Please refer to Use Case Examples for more information.

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 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, which are common in clinical and environmental studies. In this use case example, we adjust the ANI threshold and k-size to balance sensitivity and specificy, with higher values increasing stringency and refining species resolution. Using a human skin metagenomic sample, we show that these parameters markedly influence species reporting highlighting the need for careful threshold selection. For more information, refer to Low abundance samples.

Metagenomic-assembled genome (MAG) fishing: YACHT can be employed to search for specific MAGs of interest within a sample by using a single MAG as the training reference database. Applying this approach to two skin metagenomic samples shows that detection strength varies with sequencing depths and coverage. This use case example illustrates how MAG fishing with YACHT is sensitive to coverage and parameter choice, emphasizing the importance of sequencing depth when assessing MAG presence. Find further detail in MAG fishing.

Synthetic metagenomes: YACHT can assess the construction of mock or synthetic microbial communities to verify that the designed microbes are present. Using a synthetic community from the literature, we show that ANI thresholds can influence accuracy where higher ANI thresholds recover most expected genomes, while lower ones can introduce false positives further highlighitng how parameter choice—particularly ANI and minimum coverage—affect sensitivity and specificity when validating synthetic community composition. For additional information, refer to 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. C., 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., Committee, P. P., Laboratory Practices of the American Society for Microbiology, C. on, & College of American Pathologists, M. R. C. of the. (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