

Sequence analysis

MHASS: Microbiome HiFi Amplicon Sequencing Simulator

Rye Howard-Stone^{1,*} and Ion I. Măndoiu¹

¹School of Computing, University of Connecticut, CT, 06269, United States

*Corresponding author. School of Computing, University of Connecticut, 371 Fairfield Way, Storrs, CT, 06269, United States. E-mail: rye.howard-stone@uconn.edu

Associate Editor: Pier Luigi Martelli

Abstract

Summary: Microbiome HiFi Amplicon Sequence Simulator (MHASS) creates realistic synthetic PacBio HiFi amplicon sequencing datasets for microbiome studies, by integrating genome-aware abundance modeling, realistic dual-barcoding strategies, and empirically derived pass-number distributions from actual sequencing runs. MHASS generates datasets tailored for rigorous benchmarking and validation of long-read microbiome analysis workflows, including ASV clustering and taxonomic assignment.

Availability and implementation: Implemented in Python with automated dependency management, the source code for MHASS is freely available at <https://github.com/rhowardstone/MHASS> along with installation instructions. Our code is also published on Zenodo at <https://doi.org/10.5281/zenodo.17486364>. The data underlying this article are available on GitHub at https://github.com/rhowardstone/MHASS_evaluation/.

1 Introduction

Long-read sequencing, particularly PacBio's High-Fidelity Circular Consensus Sequencing (CCS) technology, has revolutionized microbiome studies by enabling the use of long amplicons for taxonomic analysis (Gehrig *et al.* 2019, Wenger *et al.* 2019). The ability to sequence multi-kilobase amplicons, such as the Titan-1TM region spanning 16S-ITS-23S (~2.5 kb), offers unprecedented taxonomic resolution compared to traditional short-read approaches (Graf *et al.* 2021). However, rigorous benchmarking of computational tools for long-read amplicon analysis remains limited due to the scarcity of realistic synthetic datasets.

Existing simulation tools often omit critical features necessary for realistic microbiome amplicon data generation. Some, such as metaSPARSim (Patuzzi *et al.* 2019) and miaSIM (Gao *et al.* 2023) produce solely an ASV count matrix while others (Ono *et al.* 2022) focus on sequence-level simulation, ignoring abundances. Some programs simulate both abundances and sequencing effects, producing full synthetic datasets, such as CAMISIM (Fritz *et al.* 2019), but do not include CCS technology. Other commonly excluded features include copy number-aware abundances accounting for multiple rRNA operons per genome, proper dual-barcode handling as used in multiplexed sequencing, and accurate pass-number distributions that directly influence CCS read accuracy. Furthermore, most simulators fail to capture the complex error profiles of PacBio HiFi reads, particularly sequence-specific patterns such as increased indel rates in homopolymer regions. As far as the authors are aware, currently there exists no tool that generates realistic multi-sample synthetic datasets for CCS amplicon sequencing of custom targets.

MHASS addresses the limitations of prior simulators by providing a modular pipeline that combines established tools with novel approaches to generate highly realistic PacBio HiFi

amplicon datasets. MHASS is the first simulator to integrate the following three components: (i) genome-aware abundance modeling, (ii) *in silico* PCR amplicon extraction, and (iii) PacBio HiFi read simulation with realistic pass-number distributions. We leveraged our recent AmpliconHunter tool (Howard-Stone and Măndoiu 2026) for genome-aware amplicon extraction to test MHASS, which uses metaSPARSim (Patuzzi *et al.* 2019) for realistic abundance modeling, PBSIM3 (Ono *et al.* 2022) for accurate subread simulation, and the PacBio CCS tool for consensus generation. A feature comparison between MHASS and existing tools is given in Table 1, available as [supplementary data](#) at *Bioinformatics* online.

2 Methods

2.1 Pipeline overview

MHASS uses a five-step pipeline to generate realistic PacBio HiFi amplicon datasets:

Genome-aware abundance simulation. MHASS utilizes metaSPARSim to generate abundance matrices at the genome level by calculating the expected abundance of each of its multiple ASVs, based on the input genome molarity as supplied by the user. This approach captures the biological reality that bacterial genomes often contain multiple, non-identical copies of the rRNA operon. Users can select from uniform, lognormal, power-law, or empirical abundance distributions based on actual microbiome profiles. The variability of an ASVs counts is given as a function of their intensity, using a negative binomial model fit to the R1 preset provided by metaSPARSim. See Fig. 1, available as [supplementary data](#) at *Bioinformatics* online for more information.

Template generation with barcodes. Each sample receives uniquely barcoded templates following PacBio's dual-barcode strategy:

A + ForwardBarcode + ForwardPrimer + ASV +
ReverseComplement(ReversePrimer) +
ReverseComplement(ReverseBarcode) + A

This structure accurately reflects the sequencing library preparation process, enabling realistic demultiplexing evaluations by allowing the simulation to include errors in barcodes.

Subread simulation. MHASS addresses a key limitation of PBSIM3 (its restriction to a single pass number per run) by implementing a novel approach based on empirical pass-number distributions. We analyzed pass-number distributions from publicly available PacBio HiFi datasets (Singer *et al.* 2016, Howard-Stone *et al.* 2025) and identified that they follow neither normal nor uniform distributions, but rather exhibit complex patterns dependent on insert length and sequencing conditions. MHASS samples from either empirical distributions extracted from real data or fitted lognormal distributions, running PBSIM3 separately for each pass number and parallelizing across CPU cores. In our validation experiments, number of passes was fit from real data (see Fig. 2, available as [supplementary data](#) at *Bioinformatics* online for details), and subread accuracy was empirically optimized for each dataset using KL divergence minimization (Fig. 3, available as [supplementary data](#) at *Bioinformatics* online). For all three datasets we retained PBSIM3's PacBio RSII default error-type difference ratio (6:55:39 for substitution: insertion: deletion), as no Sequel II-specific defaults are published.

CCS consensus. Simulated subreads undergo CCS processing using PacBio tools (v6.4.0+), yielding consensus sequences with realistic base quality profiles. The CCS algorithm is applied identically to both simulated and real data.

Data consolidation and cleanup. Final reads are combined into a single FASTQ file with proper formatting and meta-data. Read headers maintain traceability to original templates while conforming to standard PacBio formatting conventions for compatibility with downstream analyses.

2.2 Validation datasets

We validated MHASS using three datasets with known ground truth: Zymo/Titan-1 (Titan-1 amplicon defined by forward primer 5'-AGRRTTYGATYHTDGYTYAG-3' and reverse primer 5'-YCNTTCCYTYDYRGACT-3', D6300 mock community, 8 genomes with even abundances, 96 samples, 186 167 total reads), ATCC/16S [full 16S amplicon defined by primers 27F (5'-AGRGTTYGATYMTGGCTCAG-3') and 1492R (5'-RGYTACCTTGTTACGACTT-3'), MSA-1003 mock community, 20 genomes with staggered abundances, 192 samples, 2 468 174 total reads, and Phylotag/16S (full 16S amplicon also defined by primers 27F and 1492R, custom mock community, 23 genomes with highly staggered abundances, 5 samples, 113 709 total reads). The 16S datasets are publicly available (Singer *et al.* 2016); the Zymo/Titan-1 dataset was provided to the authors by Intus Biosciences. More information on these datasets is given in Tables 2–5, available as [supplementary data](#) at *Bioinformatics* online. The amplicon multiplicity patterns [“Amplitypes” (Howard-Stone and Măndoiu 2026)] for these mock communities are depicted in Figs 4–6, available as [supplementary data](#) at *Bioinformatics* online.

2.3 Evaluation metrics

We evaluated simulation realism through multiple approaches: (i) Edit Distance Distributions—comparing the distribution of minimum global edit distances between reads and reference sequences; (ii) Abundance Correlation—Pearson correlation between expected and observed ASV abundances at both genome and ASV levels; (iii) Error Profiles—analysis of substitution, insertion, and deletion patterns as a function of position in the amplicon.

3 Results and discussion

3.1 Edit distance distributions match real data

Across all three mock communities, MHASS generated reads with edit distance distributions that closely mirrored those observed in real sequencing data. Figure 1 presents per-genome violin plots of the edit distance between each read and its nearest reference ASV, with real data shown on the left and simulated data on the right.

The Zymo/Titan-1 dataset displays a distinctive bimodal distribution, most pronounced in genomes like *E. coli*, *P. aeruginosa*, and *S. enterica*, where two distinct peaks emerge. MHASS successfully captured this characteristic pattern, indicating it replicates both low-error and high-error modes of CCS read distributions for complex amplicon mixtures.

In contrast, the ATCC/16S mock community exhibited unimodal distributions across genomes, reflecting the more conserved and less variable nature of the 16S amplicon sequences. MHASS closely reproduced the shape and spread of these distributions, with consistent error profiles observed across all 20 species.

Finally, the Phylotag/16S dataset posed the most stringent challenge due to its highly staggered taxonomic abundances. Nevertheless, MHASS maintained realistic error behavior even in low-abundance genomes. Although the rarest genome (*N. dassonvillei*) was assigned no real reads, the overall shape and central tendency of most distributions remained well-aligned with real data.

3.2 Abundance correlation

Expected and observed counts were correlated for both real and simulated data: Zymo/Titan-1 (even abundances): $R^2 = 0.662$ (real), $R^2 = 0.519$ (simulated); ATCC/16S (staggered): $R^2 = 0.956$ (real), $R^2 = 0.999$ (simulated); Phylotag/16S (highly staggered): $R^2 = 0.794$ (real), $R^2 = 0.944$ (simulated). Genome-level agreement between expected and observed abundances is visualized for both real and simulated data for all datasets in Fig. 7, available as [supplementary data](#) at *Bioinformatics* online. MHASS fits a negative binomial model to predict the squared coefficient of variation of each ASV between samples based on its expected abundance using the values in metaSPARSim's R1 preset (Patuzzi *et al.* 2019), thus modeling realistic biological variability between replicates. Parameter fitting is shown in Fig. 1, available as [supplementary data](#) at *Bioinformatics* online. This choice provides moderate biological variability typical of microbiome data. The NA values frequently returned by metaSPARSim's parameter estimation procedure are emulated when the sampled variability from the negative binomial model would be <0 . When variability is NA, metaSPARSim converts it to 0, resulting in no biological variability—all replicates receive identical abundance values before the technical sampling step.

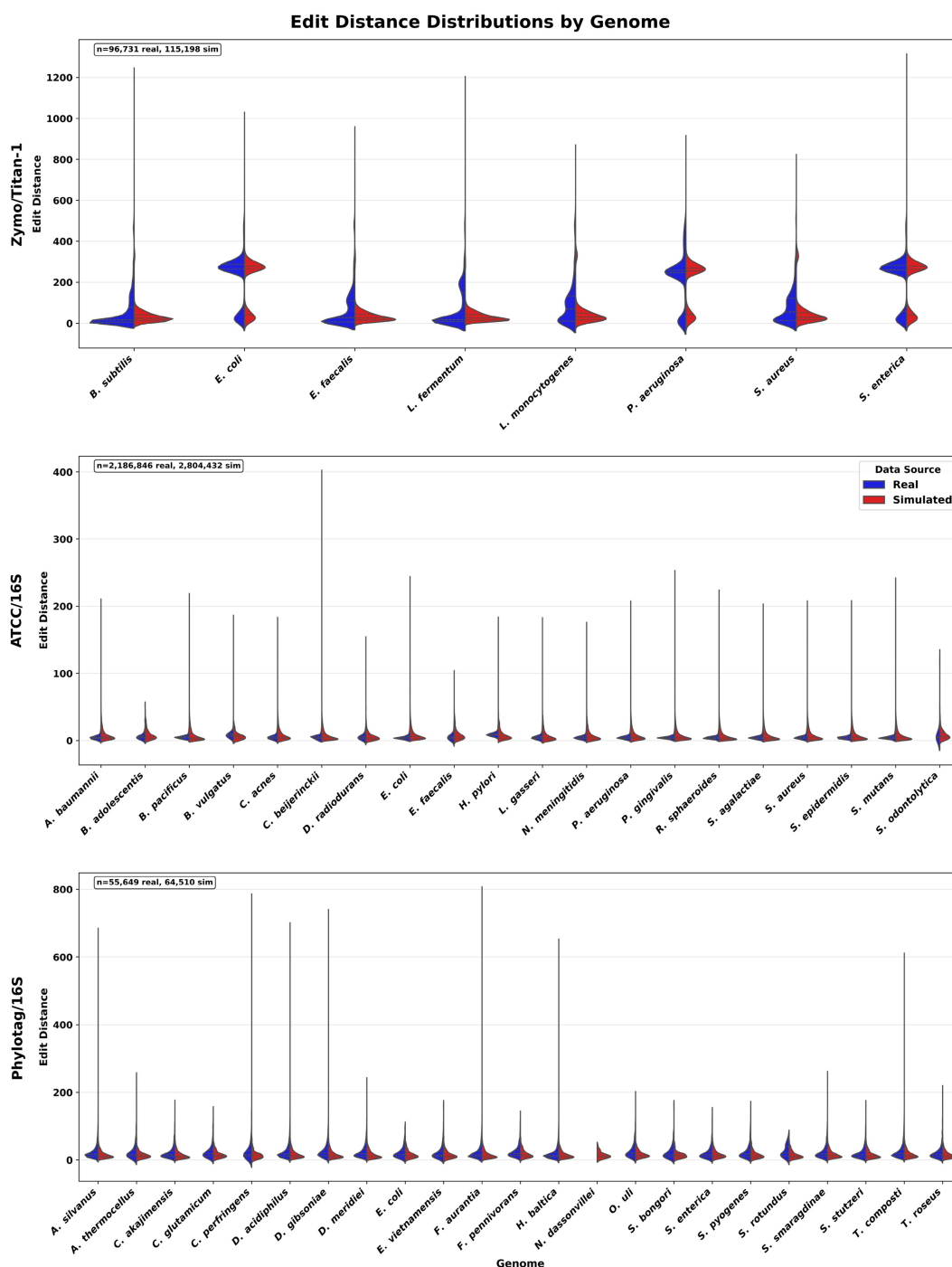


Figure 1. Edit distance distribution from read to nearest reference ASV by genome for both real and MHASS-simulated reads, with length 1400–1700 for 16S and 2000–3000 for Titan-1. Each panel corresponds to a different mock community: Zymo/Titan-1 (top), ATCC/16S (middle), and Phylotag/16S (bottom). Left violins show distributions from real CCS data; right violins show MHASS-simulated reads. Total number of reads surviving QC steps for each dataset are indicated in the top left. No misassignment of simulated reads occurred; all simulated reads were found to be closest to the original ASV sequences from which they were simulated.

3.3 Accurate modeling of error types

Analysis of error types revealed that the secondary peak in Zymo/Titan-1 edit distances resulted from insertions concentrated at the ends of the reads (positions ~2100–2500 of the alignment). Both real and simulated data showed elevated insertion rates in this region, with the proportion of insertions increasing from baseline to >90% at position ~2100 until the end of the alignment (see Fig. 8, available as [supplementary data](#) at [Bioinformatics](#) online). This pattern was absent

in 16S-only datasets, confirming its association with the Titan-1 amplicon.

4 Implementation and performance

MHASS is implemented primarily in Python 3.6+ with the following key dependencies: metaSPARSim 1.1.2 ([Patuzzi et al. 2019](#)) for abundance simulation (R 4.0+), PBSIM3 ([Ono et al. 2022](#)) for subread simulation, and PacBio CCS

tools (Wenger *et al.* 2019) 6.4.0 for consensus calling. It also requires conda (23.3.1) and Samtools (1.10) for dependency management and sequence processing. Installation is automated through a bash script that creates appropriate conda environments and downloads required models, compiling all dependencies for the target system:

```
git clone https://github.com/rhowardstone/MHASS.git
cd MHASS
bash install_dependencies.sh
conda activate mhass
```

4.1 Performance benchmarks

Performance was measured on a virtual machine configured with 180 virtual cores and 374GB RAM running on a Dell PowerEdge R7525 server with two AMD EPYC 7552 48-core CPUs. Simulation time and maximum memory usage scale linearly with the number of simulated reads, with additional overhead from the empirical pass-number sampling approach: Zymo/Titan-1 dataset (96 samples, ~2k reads/sample): 40 min; ATCC/16S dataset (192 samples, ~13k reads/sample): 6.75 h; Phylotag/16S dataset (5 samples, ~23k reads/sample): 23 min. The runtime and memory usage statistics for each dataset are summarized in Table 2, available as [supplementary data](#) at *Bioinformatics* online.

4.2 Parameter guidelines

Key parameters affecting simulation realism include: sub-read accuracy (default 0.65, based on PBSIM3 calibration, see Fig. 3, available as [supplementary data](#) at *Bioinformatics* online for more information), np distribution type (“empirical” or “lognormal”), genome distribution (choice significantly impacts abundance patterns), and np min, np max (pass number range, default: 2–59). Reproducible scripts to fit the parameters used in this evaluation are provided in our supplementary GitHub: https://github.com/rhowardstone/MHASS_evaluation.

5 Conclusion

MHASS addresses a critical need in the microbiome community for realistic long-read amplicon simulators. Our validation experiments demonstrate that MHASS accurately reproduces key characteristics of real PacBio HiFi data, including complex error patterns and abundance distributions. The presence of sequence-specific insertion patterns in Titan-1 amplicons highlights the importance of region-aware error modeling. This finding has implications for ASV calling algorithms, which may need to account for region-specific error rates to avoid oversplitting true biological variants.

Current limitations include: (i) PCR bias modeling remains simplistic, not accounting for primer-specific amplification efficiencies; (ii) Chimera formation is not explicitly modeled; (iii) Adapter dimer formation and other library preparation artifacts are not included. Future versions will address these limitations and extend support to other long-read platforms (ONT) and newer PacBio chemistries (Revio). While the current version of MHASS does not include these features, the

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics, 2025, 42, 1–4

<https://doi.org/10.1093/bioinformatics/btaf656>

Applications Note

impact on the presented benchmarking results is minimal, as we impose a minimum and maximum length when computing metrics for both real and simulated reads.

MHASS provides a much-needed tool for the microbiome research community, enabling rigorous benchmarking of analytical pipelines for PacBio HiFi amplicon sequencing. Its integration of genome-aware modeling, empirical realism, and streamlined usability makes it an essential resource for robust computational method development in microbiome research.

Author contributions

Rye Howard-Stone (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [equal], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]) and Ion Măndoiu (Conceptualization [supporting], Data curation [supporting], Formal analysis [supporting], Funding acquisition [lead], Investigation [equal], Methodology [supporting], Project administration [lead], Resources [lead], Software [supporting], Supervision [lead], Validation [supporting], Visualization [supporting], Writing—original draft [supporting], Writing—review & editing [supporting])

Supplementary data

[Supplementary data](#) is available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

None declared.

References

- Fritz A, Hofmann P, Majda S *et al.* Camisim: simulating metagenomes and microbial communities. *Microbiome* 2019;7:17. <https://doi.org/10.1186/s40168-019-0633-6>
- Gao Y, Şimşek Y, Gheysen E *et al.* miasim: an R/bioconductor package to easily simulate microbial community dynamics. *Methods Ecol Evol* 2023;14:1967–80. <https://doi.org/10.1111/2041-210X.14129>
- Gehrig JL, Venkatesh S, Chang H *et al.* Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* (1979) 2019;365:eaau4732. <https://doi.org/10.1126/science.aau4732>
- Graf J, Ledala N, Caimano MJ *et al.* High-resolution differentiation of enteric bacteria in premature infant fecal microbiomes using a novel rRNA amplicon. *mBio* 2021;12:e03656-20. <https://doi.org/10.1128/mBio.03656-20>
- Howard-Stone R, Gerwin P, Capunitan D *et al.* Cecal microbiome transplantation without antibiotic preconditioning standardizes murine microbiomes. *Front Microbiol* 2025;16:1632210. <https://doi.org/10.3389/fmicb.2025.1632210>
- Howard-Stone R, Măndoiu IL. Ampliconhunter: a scalable tool for per amplicon prediction from microbiome samples. In: *Computational Advances in Bio and Medical Sciences (ICCABS 2025), volume 15599 of Lecture Notes in Computer Science*, Cham: Springer, 2026, 329–44. https://doi.org/10.1007/978-3-032-02489-3_25
- Ono Y, Hamada M, Asai K. PBSIM3: a simulator for all types of PACBIO and ONT long reads. *NAR Genom Bioinform* 2022;4:lqac092. <https://doi.org/10.1093/nargab/lqac092>