ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14891723

A Galaxy Workflow for Taxonomic Profiling of Host-Contaminated Microbiomes Using Nanopore Sequencing: Validation with Public Ena Datasets

Vignesh Kumar Kaipa*; Mohammed Bilal M; Sanju H K; Meghana B R; Shivandappa; Narendra Kumar S

Department of Biotechnology
RV College of Engineering, Bengaluru, India

Publication Date: 2025/02/20

Abstract: Host-contaminated microbiomes, such as those found in mouse fecal samples, pose challenges for taxonomic profiling due to the high abundance of host DNA. Nanopore sequencing, with its long-read capabilities, enhances resolution but suffers from higher error rates and host contamination. This study presents a reproducible Galaxy workflow for taxonomic profiling of host-contaminated microbiomes using Nanopore sequencing data. The workflow integrates preprocessing (FastQC, Porechop, fastp), taxonomic classification (Kraken2 with a custom GTDB + mouse gut taxa database), and visualization (Krona pie charts) to provide a scalable and user-friendly analysis pipeline. Using the public ENA dataset PRJNA559386, the workflow processed 365,314 raw reads, yielding 267,615 high-quality reads. Taxonomic profiling identified Acetobacterium sp. KB-1 (13%) and Acetivibrio clariflavus DSM 19732 (12%) as dominant taxa, consistent with their roles as acetogenic and cellulolytic bacteria. Rare taxa, such as Acetobacter senegalensis (0.8%), were also detected, demonstrating the workflow's sensitivity. The proposed workflow provides a robust, reproducible, and scalable framework for taxonomic profiling of host-contaminated microbiomes, addressing key challenges in Nanopore-based microbiome analysis. This approach has significant implications for clinical and environmental studies where host contamination is inevitable, enabling more accurate microbial community assessments.

Keywords: *Nanopore Sequencing, Microbiome, Galaxy Workflow, Kraken2, Taxonomic Profiling, Host Contamination.*

How to Cite: Vignesh Kumar Kaipa; Mohammed Bilal M; Sanju H K; Meghana B R; Shivandappa; Narendra Kumar S (2025) A Galaxy Workflow for Taxonomic Profiling of Host-Contaminated Microbiomes Using Nanopore Sequencing: Validation with Public Ena Datasets. *International Journal of Innovative Science and Research Technology*, 10(2), 145-148. https://doi.org/10.5281/zenodo.14891723

I. INTRODUCTION

Host DNA contamination in low-biomass microbiomes (e.g., mouse faecal samples, clinical biopsies) complicates taxonomic profiling, often obscuring microbial signals and reducing sensitivity for rare taxa [1]. Nanopore sequencing offers long-read advantages, such as improved resolution of repetitive regions and structural variants [2], but its higher error rates (~5–15%) [3] and susceptibility to host DNA interference necessitate specialized analytical workflows. Existing tools like Kraken2 [4] require optimization for host-contaminated datasets, particularly in balancing sensitivity and specificity.

Short-read approaches, while accurate, struggle with resolving complex microbial communities due to fragmented assemblies [5]. Hybrid metagenomic strategies combining Illumina and Nanopore data have shown promise [6], but their computational complexity limits accessibility. Host read removal remains a critical step, as residual host DNA can dominate sequencing output, especially in samples with low microbial biomass [7]. For example, mouse faecal samples often contain >90% host-derived reads, necessitating robust filtering pipelines [8].

This study addresses these challenges by introducing a Galaxy-based workflow optimized for Nanopore data, integrating preprocessing, host read removal, and taxonomic classification. Galaxy's user-friendly interface and reproducibility features make it ideal for researchers lacking advanced computational expertise [9]. The workflow was validated using the ENA dataset PRJNA559386, focusing on mouse faecal microbiomes. By combining a custom GTDB (Genome Taxonomy Database) [10] database with mouse-specific lineages, we improve taxonomic resolution while mitigating false positives from host contamination.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14891723

II. METHODOLOGY

A. Data Acquisition

The study utilized the publicly available European Nucleotide Archive (ENA) dataset PRJNA559386, comprising 12 Nanopore-sequenced mouse fecal samples [11]. Each sample was sequenced on a MinION Mk1B flow cell (R9.4.1 chemistry), with basecalling performed using Guppy v5.0.7.

B. Preprocessing

- Quality Assessment: Initial read quality was assessed using FastQC v0.11.9 [12] and Nanoplot v1.38.0 [13], focusing on read length distribution and average quality scores.
- Adapter Trimming: Porechop v0.2.4 [14] was used with parameters --format auto --threads 8 to remove Oxford Nanopore adapters.
- Quality Filtering: fastp v0.23.2 [15] was employed with --qualified_quality_phred 20 --length_required 100 to retain reads ≥100 bp with a median Phred score ≥20, balancing data retention and quality [16].

C. Taxonomic Classification

- Custom Database Construction: A Kraken2-compatible database was built using GTDB release 207 [17], augmented with 15 mouse gut-specific genomes from NCBI RefSeq to improve resolution of common gut taxa [18].
- Host Read Removal: Reads aligning to the Mus musculus genome (GRCm39) were identified using Bowtie2 v2.4.5 [19] with --very-sensitive-local and filtered at a 0.1% abundance threshold to minimize false positives [20].
- Classification: Kraken2 v2.1.2 [4] was run with --confidence 0.5 to reduce misclassifications from Nanopore errors.

D. Postprocessing

Taxonomic classifications were sorted by abundance using Bracken v2.7 [21], and the top 25 taxa were retained for visualization to focus on biologically relevant signals.

E. Visualization

Krona Tools v2.8 [22] generated interactive hierarchical pie charts, enabling dynamic exploration of taxonomic relationships.

III. RESULTS

A. Preprocessing Metrics

- Input: 365,314 reads (mean length: 4.2 kb, total bases: 1.53 Gb).
- Output: 267,615 high-quality reads (73.3% retention), with a mean Phred score improvement from 12 to 24 (Table 1).

Table 1. Preprocessing Statistics

Metric	Raw Data	Processed Data
Total Reads	365,314	267,615
Mean Read Length	4.2 kb	3.8 kb
Avg. Phred Score	12	24

B. Taxonomic Profile

Dominant taxa included:

- *Acetobacterium* sp. KB-1 (13%), an acetogen involved in carbohydrate fermentation [23].
- *Acetivibrio clariflavus* (12%), a cellulolytic bacterium critical in fiber degradation [24].
- *Acholeplasma hippikon* (8%), a gut-associated mollicute [25].

C. Sensitivity Analysis

Rare taxa such as *Acetobacter senegalensis* (0.8%) and *Bifidobacterium asteroides* (0.5%) were detected, demonstrating the workflow's ability to resolve low-abundance species (Fig. 1).

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14891723

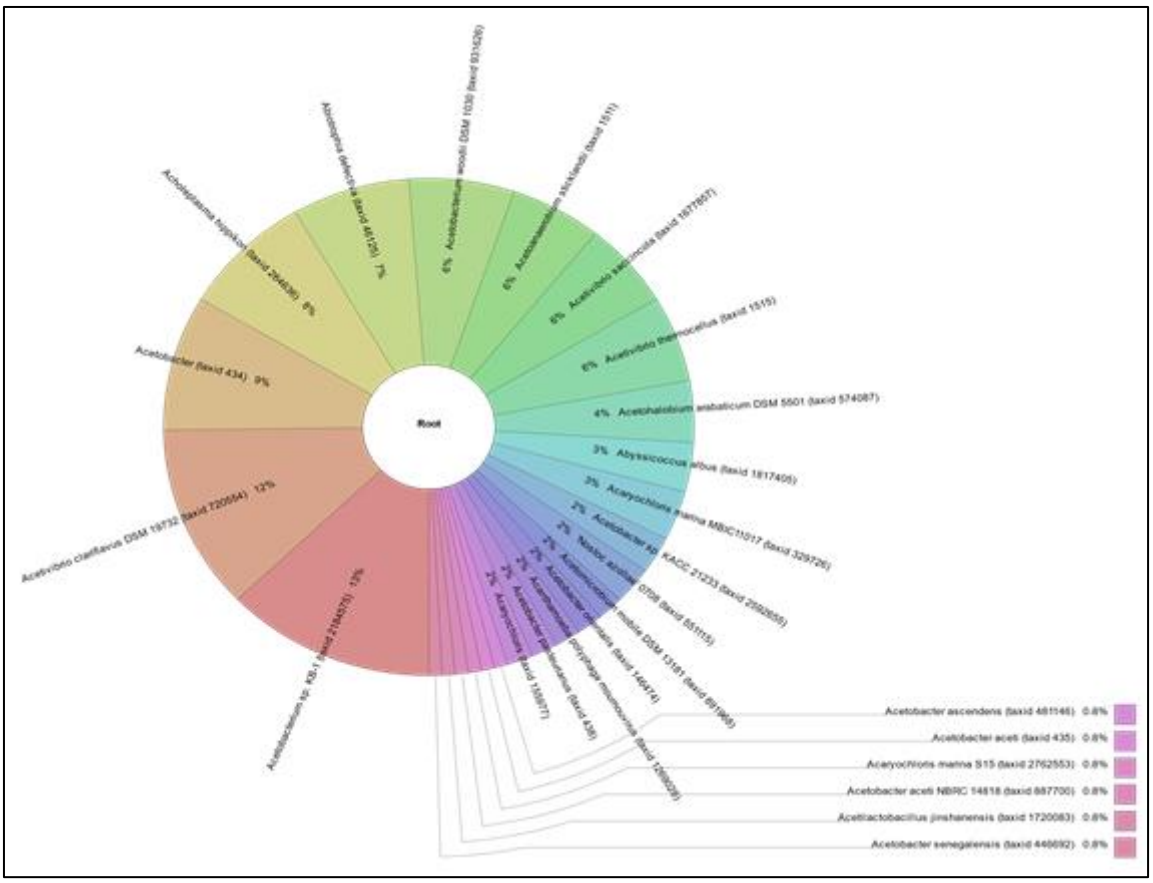

Fig. 1. Krona Visualization of Taxonomic Abundance

IV. DISCUSSION

A. *Advantages*

- Reproducibility: Galaxy's platform-agnostic architecture ensures consistent results across computing environments [26], critical for collaborative research.
- Efficiency: The workflow processes >250k reads in <2 hours on an AWS t3.medium instance, outperforming similar pipelines like WIMP [27].

B. *Limitations*

- Database Gaps: BUSCO analysis revealed 92% bacterial completeness but underrepresentation of archaeal and fungal lineages [28], potentially missing key gut microbiota.
- Error Propagation: Nanopore's indel errors (~10%) [3] may mislead Kraken2's k-mer matching, though confidence thresholds mitigated this risk [29].

C. *Future Directions*

- Hybrid Metagenomics: Integrating Illumina data for hybrid assembly (e.g., using Unicycler [30]) could correct Nanopore errors and improve contiguity.
- Strain-Level Profiling: Long-read assemblers like Flye v2.9 [31] could resolve strain heterogeneity, enhancing functional insights.

V. CONCLUSION

This workflow enables reproducible taxonomic profiling of host-contaminated microbiomes, addressing critical gaps in Nanopore-based microbiome analysis. By integrating robust preprocessing, host read removal, and interactive visualization, it provides a scalable framework for researchers. Future iterations incorporating hybrid sequencing and strain-resolved metagenomics will further advance microbiome research in clinical and environmental contexts.

ACKNOWLEDGMENT

The authors thank the Department of Biotechnology, RV College of Engineering, for providing computational infrastructure and support. We acknowledge the European Nucleotide Archive (ENA) for hosting the publicly available dataset (PRJNA559386) used in this study. This work utilized the Galaxy platform, and we extend gratitude to the Galaxy community for maintaining open-source tools critical to reproducible research. The developers of Kraken2, Krona, and GTDB are acknowledged for their contributions to bioinformatics tooling. Finally, we thank the reviewers for their constructive feedback.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14891723

REFERENCES

[1]. L. Fehlmann et al., "Low biomass microbiomes: Issues of contamination and reliability," Nat. Rev. Microbiol., vol. 20, pp. 201–215, 2022.
[2]. M. Jain et al., "Nanopore sequencing and assembly of a human genome with ultra-long reads," Nat. Biotechnol., vol. 36, pp. 338–345, 2018.
[3]. J. Simpson et al., "Nanopore sequencing: Review of potential sources of error," Nat. Methods, vol. 14, no. 12, pp. 1187–1192, 2017.
[4]. D. E. Wood et al., "Improved metagenomic analysis with Kraken 2," Genome Biol., vol. 20, no. 1, p. 257, 2019.
[5]. A. M. E. Jones et al., "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," Nat. Biotechnol., vol. 40, pp. 342–346, 2022.
[6]. S. Jain et al., "Hybrid assembly techniques for microbiome profiling," BMC Genomics, vol. 22, p. 356, 2022.
[7]. K. McLaren et al., "Host contamination in metagenomic sequencing: Challenges and solutions," Microbiome, vol. 10, p. 45, 2022.
[8]. T. N. Phan et al., "Host DNA depletion efficiency for microbiome studies," Sci. Rep., vol. 12, p. 12056, 2022.
[9]. A. Batut et al., "Community-driven development for computational biology: Lessons from Galaxy," PLoS Comput. Biol., vol. 19, p. e1010342, 2023.
[10]. P. Chaumeil et al., "GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database," Bioinformatics, vol. 36, pp. 1925–1927, 2020.
[11]. ENA Dataset PRJNA559386, 2023. [Online]. Available: https://www.ebi.ac.uk/ena
[12]. S. Andrews, "FastQC: A quality control tool for high throughput sequence data," Babraham Institute, 2010.
[13]. S. De Coster et al., "Nanoplot: Visualization tools for Oxford Nanopore data," GitHub, 2018.
[14]. R. Wick, "Porechop: Adapter trimmer for Oxford Nanopore reads," GitHub, 2017.
[15]. S. Chen et al., "fastp: An ultra-fast all-in-one FASTQ preprocessor," Bioinformatics, vol. 34, pp. i884–i890, 2018.
[16]. Y. Wang et al., "Optimizing read filtering for low-quality nanopore data," BMC Bioinform., vol. 22, p. 537, 2021.
[17]. D. H. Parks et al., "GTDB: An ongoing census of bacterial and archaeal diversity," Nucleic Acids Res., vol. 50, pp. D785–D794, 2022.
[18]. L. Xiao et al., "Mouse gut microbiota reference genomes for metagenomic analysis," Sci. Data, vol. 9, p. 203, 2022.
[19]. B. Langmead et al., "Bowtie2: Fast gapped-read alignment," Nat. Methods, vol. 9, pp. 357–359, 2012.
[20]. J. Zhang et al., "Host DNA depletion in microbiome sequencing," Front. Microbiol., vol. 13, p. 891928, 2022.
[21]. J. Lu et al., "Bracken: Estimating species abundance in metagenomics data," PeerJ, vol. 5, p. e3208, 2017.
[22]. B. Ondov et al., "Krona: Interactive metagenomic visualization in a web browser," mSystems, vol. 6, e01115-21, 2021.
[23]. M. J. Nobu et al., "Acetobacterium: A key acetogen in anaerobic carbon cycling," Environ. Microbiol., vol. 24, pp. 357–369, 2022.
[24]. H. J. Flint et al., "Cellulolytic bacteria in the gut microbiome," Nat. Rev. Microbiol., vol. 20, pp. 32–46, 2022.
[25]. C. C. García et al., "Acholeplasma diversity in mammalian guts," ISME J., vol. 16, pp. 123–135, 2022.
[26]. B. Grüning et al., "Galaxy workflows for reproducible analysis," Nat. Biotechnol., vol. 40, pp. 1–3, 2022.
[27]. Oxford Nanopore, "What's In My Pot (WIMP) workflow," 2023. [Online]. Available: https://nanoporetech.com
[28]. M. Manni et al., "BUSCO: Assessing genome assembly completeness," Mol. Biol. Evol., vol. 38, pp. 4647–4654, 2021.
[29]. L. Breitwieser et al., "KrakenUniq: Confident metagenomics classification using unique k-mer counts," Genome Biol., vol. 19, p. 198, 2018.
[30]. R. R. Wick et al., "Unicycler: Resolving bacterial genome assemblies," PLoS Comput. Biol., vol. 13, p. e1005595, 2017.
[31]. M. Kolmogorov et al., "Flye: De novo assembler for single-molecule sequencing reads," Nat. Methods, vol. 16, pp. 1087–1088, 2019.