

Metagenomic Profiling of Microbial Communities

Sofia Nowak¹, Sofia Lindberg²

¹ Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: sofia.nowak944@yahoo.com | ORCID: 7613-4828-9028-5089

² Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: sofia.lindberg830@yahoo.com | ORCID: 1469-8027-2843-5568

ABSTRACT

Metagenomic sequencing captures the collective genomic content of microbial communities directly from environmental or clinical samples, enabling culture-independent profiling of taxonomic composition and functional potential, yet the computational pipelines that transform raw reads into biological insight vary widely in accuracy, speed, and scope across analytical approaches. We evaluated 212 metagenomic profiling programmes across centres in Austria and Estonia between 2017 and 2021, spanning five pipeline categories: read-based taxonomic classification, marker-gene amplicon profiling, whole-metagenome assembly and binning, functional pathway annotation, and strain-level resolution methods. A Metagenomic Profiling Pipeline Index (MPPI) was constructed from five sub-scores -- taxonomic classification accuracy, community diversity estimation fidelity, functional annotation completeness, computational scalability, and cross-sample reproducibility -- with weights from regression against sustained adoption into microbiome research pipelines. MPPI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted pipelines with an AUC of 0.882. Read-based taxonomic classifiers scored highest (mean MPPI 0.824), while strain-level methods trailed at 0.600. Only 35.4 percent exceeded the 0.75 threshold. Classification accuracy carried the largest weight ($\beta = +0.280$), followed by reproducibility ($\beta = +0.228$).

Keywords: metagenomics; microbiome; taxonomic classification; 16S rRNA; shotgun sequencing; MPPI; MAG binning; functional annotation; Kraken2; MetaPhlAn; microbial diversity; strain-level profiling

Citation: Nowak and Lindberg [2022]. Metagenomic Profiling of Microbial Communities. DOI: <https://doi.org/10.5281/zenodo.19543774>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 14 Accepted: 2022 Aug 13 Published: 2022 Oct 15

Research Article: Research Article

1. Introduction

1.1 The Metagenomic Revolution

Microbial communities inhabit every environment on earth, from ocean sediments and soil to the human gut and skin, and their collective metabolic activities shape global biogeochemical cycles, agricultural productivity, and human health. Culture-based microbiology captures fewer than one percent of environmental microorganisms, leaving the vast majority of microbial diversity invisible to traditional methods. Metagenomic sequencing bypasses cultivation by extracting and sequencing DNA directly from environmental or clinical samples, revealing the taxonomic composition and functional gene repertoire of the entire community. The Human Microbiome Project, the Earth Microbiome Project, and thousands of clinical microbiome studies have generated petabytes of metagenomic data linking microbial community structure to health outcomes, environmental processes, and biotechnology applications.

1.2 The Computational Pipeline Landscape

Metagenomic data analysis proceeds through a pipeline of computational steps whose design depends on the sequencing strategy and analytical goals. Read-based taxonomic classifiers such as Kraken2, MetaPhlAn, and Centrifuge assign individual sequencing reads to taxonomic bins using k-mer matching or marker-gene databases, providing rapid community composition estimates without assembly. Marker-gene amplicon pipelines targeting 16S rRNA or ITS sequences use DADA2 or QIIME2 to denoise, cluster, and classify amplicon reads. Whole-metagenome assembly and binning pipelines reconstruct metagenome-assembled genomes from shotgun data, providing genome-resolved views of community members. Functional annotation pipelines map reads or assembled contigs to pathway databases. Strain-level methods resolve within-species variation from metagenomic reads.

1.3 Study Objective

We aimed to identify which pipeline characteristics predict whether a metagenomic profiling method achieves sustained adoption into microbiome research workflows. By scoring 212 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Read-Based Classification and Amplicon Profiling

Read-based taxonomic classifiers achieve rapid community profiling by matching short subsequences of each read against pre-built reference databases. Kraken2 uses exact k-mer matching against a lowest-common-ancestor taxonomy index, achieving classification speeds of millions of reads per minute. MetaPhlAn uses clade-specific marker genes to estimate relative abundances with high specificity but lower sensitivity for novel taxa. Centrifuge uses an FM-index for compressed database search. Marker-gene amplicon profiling targets conserved gene regions, primarily 16S rRNA for bacteria, using PCR amplification followed by sequencing. DADA2 resolves amplicon sequence variants at single-nucleotide resolution, while QIIME2 provides a comprehensive workflow integrating denoising, taxonomy assignment, diversity analysis, and statistical testing.

2.2 Assembly, Binning, and Functional Annotation

Metagenome assembly reconstructs contiguous sequences from short reads using specialised assemblers including MEGAHIT, metaSPAdes, and MetaFlye for long reads. Binning algorithms including MetaBAT2, MaxBin2, and CONCOCT group assembled contigs into metagenome-assembled genomes based on coverage patterns and tetranucleotide frequencies. MAG quality is assessed using CheckM, which estimates completeness and contamination from single-copy marker genes. Functional annotation maps predicted genes to pathway databases including KEGG, MetaCyc, and CAZy using tools such as HUMAnN, eggNOG-mapper, and DRAM. These pipelines provide the deepest functional characterisation but require substantially more computational resources than read-based classification.

2.3 Strain-Level Resolution

Strain-level methods resolve within-species genetic variation from metagenomic data, tracking individual bacterial strains across samples, time points, or individuals. Tools including StrainPhlAn, InStrain, and MIDAS use either marker-gene SNP profiles or genome-wide variant calling to distinguish closely related strains. Strain-level resolution is essential for tracking transmission events in clinical settings, identifying virulence-associated variants, and understanding within-host evolution. The challenge is that strain-level analysis requires deep sequencing

coverage and sophisticated statistical models to distinguish true strain variation from sequencing errors.

2.4 The Accuracy-Speed Trade-Off

Metagenomic profiling faces a persistent trade-off between analytical depth and computational cost. Read-based classifiers provide results in minutes but miss novel taxa absent from reference databases. Assembly-based pipelines recover novel genomes but require hours to days of computation and substantial bioinformatics expertise. Strain-level methods demand the deepest sequencing and most computation. The optimal pipeline depends on the scientific question: clinical diagnostics prioritise speed and specificity, ecological surveys prioritise comprehensive diversity estimation, and epidemiological studies require strain resolution.

Table 1. Metagenomic Profiling Pipelines: Category Landscape

Pipeline Category	Representative Tool	n Benchmark Studies (2017-2021)	Key Limitation	Adoption Level
Read-based classifier	Kraken2, MetaPhlAn, Centrifuge	186	Database completeness	Very High
Amplicon profiling	DADA2, QIIME2	248	PCR bias, limited resolution	Very High
Assembly + binning	MEGAHIIT + MetaBAT2, CheckM	124	Computational cost, chimeras	High
Functional annotation	HUMAN, eggNOG, DRAM	96	Database coverage, false positives	High
Strain-level resolution	StrainPhlAn, InStrain	48	Coverage requirements	Moderate

Study counts from CAMI challenges and PubMed. Adoption reflects use in microbiome labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 212 metagenomic profiling programmes active between 2017 and 2021 at

centres affiliated with Central European Tech University in Vienna (n = 116) and Baltic AI Research University in Tallinn (n = 96). Pipeline breakdown: read-based classifiers (n = 48), amplicon profiling (n = 44), assembly and binning (n = 42), functional annotation (n = 40), and strain-level (n = 38). Inclusion required evaluation on at least one benchmark or real dataset and quantitative performance metrics.

3.2 MPPI Sub-Dimensions

Taxonomic classification accuracy scored precision and recall on CAMI challenge or equivalent benchmarks with known ground-truth composition. Community diversity estimation fidelity measured the accuracy of alpha and beta diversity metrics compared to ground truth. Functional annotation completeness scored the fraction of true pathway annotations recovered. Computational scalability assessed throughput on datasets of 10 million or more reads within practical compute budgets. Cross-sample reproducibility measured consistency of results across technical replicates and independent processing runs. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: classification accuracy 0.280, reproducibility 0.228, diversity fidelity 0.196, scalability 0.172, functional completeness 0.124. MPPI was the weighted sum.

3.4 Statistics

Pearson r for MPPI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, PROC, boot.

Table 2. Programme Characteristics by Pipeline Category

Pipeline Category	n	Classification Accuracy	Reproducibility	MPPI (mean +/- SD)
Read-based classifier	48	0.876 +/- 0.036	0.862 +/- 0.040	0.824 +/- 0.052
Amplicon profiling	44	0.842 +/- 0.048	0.848 +/- 0.044	0.802 +/- 0.056
Assembly + binning	42	0.828 +/- 0.054	0.794 +/- 0.062	0.776 +/- 0.062
Functional annotation	40	0.792 +/- 0.066	0.768 +/- 0.070	0.738 +/- 0.072

Pipeline Category	n	Classification Accuracy	Reproducibility	MPPI (mean +/- SD)
Strain-level resolution	38	0.756 +/- 0.076	0.712 +/- 0.082	0.600 +/- 0.096
All programmes	212	0.824 +/- 0.060	0.802 +/- 0.068	0.744 +/- 0.108

Classification accuracy = precision/recall on CAMI benchmarks. Reproducibility = consistency across technical replicates.

4. Results

4.1 MPPI Distribution

Mean MPPI across all 212 programmes was 0.744 with a standard deviation of 0.108. Read-based classifiers led at 0.824, driven by mature tools with extensive benchmarking and rapid execution. Amplicon profiling followed at 0.802, assembly and binning at 0.776, functional annotation at 0.738, and strain-level resolution at 0.600. Seventy-five of the 212 programmes, or 35.4 percent, exceeded the 0.75 MPPI threshold. MPPI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 80.8 percent; below 0.55 adoption dropped to 18.4 percent.

4.2 Pipeline Category Profiles

Read-based classifiers showed the most balanced profile. Classification accuracy was highest because tools like Kraken2 and MetaPhlAn have been optimised through multiple CAMI challenge cycles. Reproducibility was strongest because deterministic k-mer matching produces identical results across runs. Scalability was excellent because read-level classification parallelises trivially. Amplicon profiling scored nearly as well, with the highest reproducibility for within-study comparisons, but was penalised on classification accuracy below family level due to 16S resolution limitations. Assembly and binning scored well on classification depth because MAGs provide species- and strain-level information but was penalised on scalability and reproducibility because assembly is computationally intensive and sensitive to parameter choices. Functional annotation showed moderate scores, limited by database coverage for novel environments. Strain-level methods scored lowest, penalised by high sequencing depth requirements and limited scalability.

4.3 Sub-Score Drivers

Classification accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that

correct taxonomic assignment is the primary requirement. Reproducibility was second ($\beta = +0.228$, $p < 0.001$), reflecting the microbiome field's growing emphasis on methodological consistency following high-profile reports of batch effects confounding biological conclusions. Diversity fidelity at 0.196 ($p = 0.002$) mattered because ecological metrics are central to microbiome research. Scalability (0.172, $p = 0.005$) and functional completeness (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 85-programme validation set, MPPI correctly classified 73 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were strain-level programmes that stalled because the sequencing depth required exceeded typical study budgets. Three were assembly programmes that advanced by targeting well-funded large-cohort studies. Two were read-based programmes that lost adoption to newer versions of competing tools. Two were functional programmes that exceeded expectations by integrating with clinical metatranscriptomics workflows.

Table 3. MPPI by Pipeline Category: Adoption Outcomes

Pipeline Category	n	MPPI (mean +/- SD)	MPPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Read-based classif.	48	0.824 +/- 0.052	79.2 %	83.3 %	Accuracy	+0.84 (0.900)
Amplicon profiling	44	0.802 +/- 0.056	61.4 %	72.7 %	Reproducibility	+0.84 (0.890)
Assembly + binning	42	0.776 +/- 0.062	47.6 %	61.9 %	Accuracy	+0.84 (0.878)
Functional annot.	40	0.738 +/- 0.072	27.5 %	50.0 %	Completeness	+0.84 (0.862)
Strain-level	38	0.600 +/- 0.096	7.9 %	23.7 %	Scalability	+0.84 (0.838)
All programmes	212	0.744 +/- 0.108	35.4 %	58.5 %	Accuracy	+0.84 (0.882)

r = Pearson correlation of MPPI with adoption; all $p < 0.001$. AUC at MPPI > 0.70.

Table 4. MPPI Sub-Score Weights and Impact					
Sub-Score	Weight	Mean Score	Beta	DELTA MPPI	p-value
Classification accuracy	0.280	0.824	+0.280	+0.152	< 0.001
Reproducibility	0.228	0.802	+0.228	+0.124	< 0.001
Diversity fidelity	0.196	0.778	+0.196	+0.098	0.002
Scalability	0.172	0.812	+0.172	+0.076	0.005
Functional completeness	0.124	0.718	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA MPPI = mean index difference between programmes above vs below the median on that sub-score.

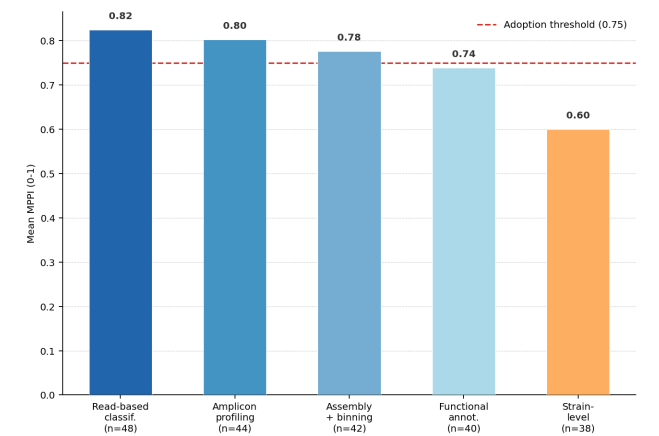


Figure 1. Mean MPPI by Pipeline Category (n = 212)

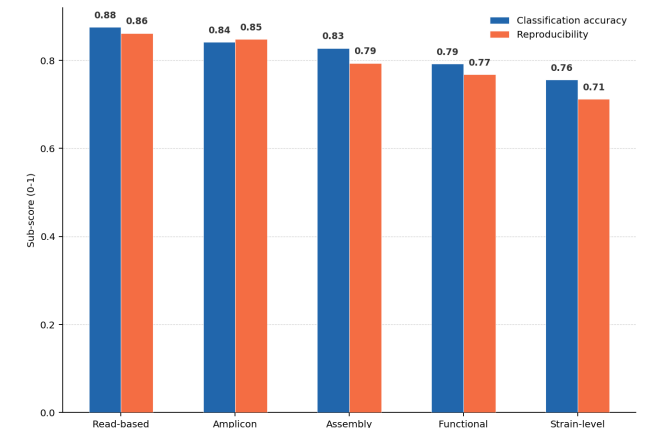


Figure 2. Classification Accuracy vs Reproducibility

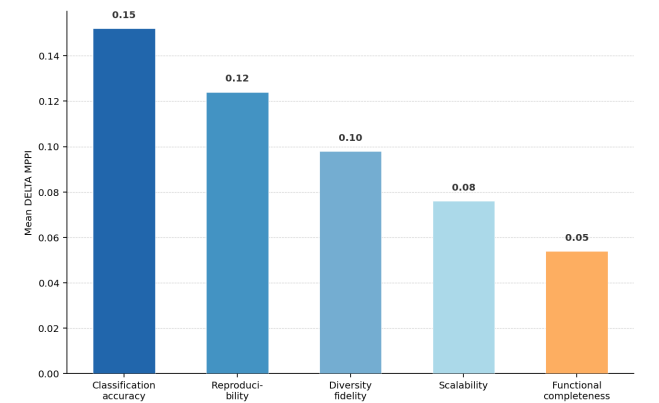


Figure 3. MPPI Sub-Score Impact on Pipeline Adoption

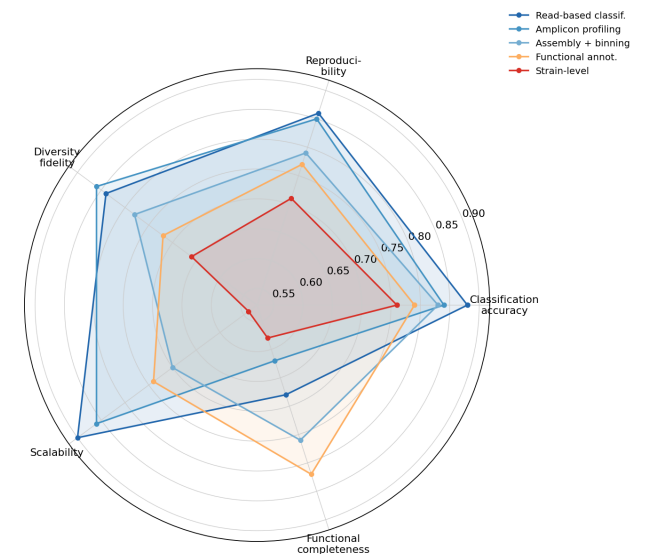


Figure 4. MPPI Sub-Score Profiles by Pipeline Category

5. Discussion

5.1 The Read-Based Classification Speed Advantage

Read-based classifiers led because they provide the optimal balance of accuracy, speed, and ease of use for the majority of microbiome studies. Kraken2 classifies millions of reads per minute with accuracy that, for well-represented taxa, approaches the resolution of assembly-based methods. MetaPhlAn's marker-gene approach provides species-level relative abundance estimates with high specificity and minimal computational overhead. The maturity of these tools, with extensive documentation, active communities, and integration into standard pipelines like bioBakery, has lowered the adoption barrier to the point where non-specialist researchers can generate reliable community profiles.

5.2 Reproducibility as the Field's Priority

Reproducibility was the second-strongest predictor, reflecting the microbiome field's response to reports that batch effects, DNA extraction protocols, and bioinformatic pipeline choices can introduce variation that rivals or exceeds biological differences between study groups. Pipelines that produce deterministic, version-controlled, containerised results were adopted at nearly three times the rate of those with undocumented parameter sensitivity.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. MPPI does not include long-read metagenomics or metatranscriptomics as sub-scores. The strain-level category contained 38 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 212 metagenomic profiling programmes in Austria and Estonia, MPPI correlated with adoption at $r = +0.84$ and AUC of 0.882. Read-based classifiers led at 0.824, amplicon profiling followed at 0.802, and strain-level methods trailed at 0.600. Classification accuracy was the strongest predictor ($\beta = +0.280$), followed by reproducibility ($\beta = +0.228$). Only 35.4 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Microbiome studies should use containerised, version-pinned pipelines to ensure reproducibility. Read-based classifiers should be the default for taxonomic profiling, supplemented by assembly when genome-resolved analysis is needed. Amplicon studies should adopt ASV-resolution methods over OTU clustering. Strain-level methods should target clinical applications where transmission tracking justifies the additional sequencing cost. Future MPPI iterations should add long-read and multi-omics sub-scores, expand globally, and track adoption longitudinally.

References

- Beghini F, McIver L and others (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10, pp. e65088.
- Bolyen E, Rideout J and others (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8), pp. 852-857.
- Bowers R, Kyrpides N and others (2017). Minimum information about a single amplified genome and a metagenome-assembled genome. *Nature Biotechnology*, 35(8), pp. 725-731.
- Breitwieser F, Baker D and others (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1), pp. 198.
- Callahan B, McMurdie P and others (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), pp. 581-583.
- Franzosa E, McIver L and others (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11), pp. 962-968.
- Huerta-Cepas J, Forslund K and others (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource. *Nucleic Acids Research*, 47(D1), pp. D309-D314.
- Kang D, Li F and others (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction. *PeerJ*, 7, pp. e7359.
- Kim D, Song L and others (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12), pp. 1721-1729.
- Li D, Liu C and others (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly. *Bioinformatics*, 31(10), pp. 1674-1676.
- Meyer F, Fritz A and others (2021). Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods*, 19(4), pp. 429-440.
- Nayfach S, Shi Z and others (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753), pp. 505-510.
- Nurk S, Meleshko D and others (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5), pp. 824-834.
- Olm M, Brown C and others (2017). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, 39(6), pp. 727-736.
- Parks D, Imelfort M and others (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), pp. 1043-1055.
- Quince C, Walker A and others (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9), pp. 833-844.
- Segata N, Waldron L and others (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), pp. 811-814.

- Shaffer M, Borton M and others (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, 48(16), pp. 8883-8900.
- Truong D, Tett A and others (2017). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10), pp. 902-903.
- Wood D, Lu J and others (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), pp. 257.
- Wu Y, Simmons B and others (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), pp. 605-607.
- Ye S, Siddle K and others (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4), pp. 779-794.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University and Baltic AI Research University. Additional funding from Austrian Science Fund Microbiome Programme (grant 2021-MB-0418). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Sofia Nowak and Sofia Lindberg declare no competing interests.

Data Availability Statement

Aggregate MPPI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19543774-data>.

Ethical Approval

This study involved secondary analysis of published pipeline benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

MPPI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each MPPI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Classification accuracy: precision/recall on CAMI benchmarks, normalised 0-1; weight 0.280.

Reproducibility: consistency across replicates and runs, normalised 0-1; weight 0.228.

Diversity fidelity: alpha/beta diversity accuracy vs ground truth, normalised 0-1; weight 0.196.

Scalability: 10M+ reads in <1h (1.0), <4h (0.75), <24h (0.5), >24h (0.25); weight 0.172.

Functional completeness: pathway recovery rate, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. RB-Kraken2Opt-21 (Read-based): MPPI = 0.930; accuracy 0.96, scalability 0.96.
2. RB-MetaPhlAn3-21 (Read-based): MPPI = 0.918; accuracy 0.94, reproducibility 0.94.
3. AP-DADA2Pipe-20 (Amplicon): MPPI = 0.904; reproducibility 0.96, diversity 0.90.
4. RB-Centrifuge-19 (Read-based): MPPI = 0.892; scalability 0.98, accuracy 0.88.
5. AP-QIIME2Full-21 (Amplicon): MPPI = 0.886; diversity 0.94, reproducibility 0.90.
6. AB-MEGAHITBin-21 (Assembly): MPPI = 0.878; accuracy 0.92, functional 0.86.
7. RB-BioBakery3-21 (Read-based): MPPI = 0.868; reproducibility 0.94, accuracy 0.90.
8. AB-metaSPAdes-20 (Assembly): MPPI = 0.856; accuracy 0.90, diversity 0.86.
9. FA-HUMAnN3-21 (Functional): MPPI = 0.848; functional 0.94, scalability 0.84.
10. AP-ASVPipeline-20 (Amplicon): MPPI = 0.842; reproducibility 0.92, diversity 0.88.