Scientific Reports

https://doi.org/10.1038/s41598-026-49725-3

Article in Press

Benchmarking short- and long-read sequencing technologies for metagenomic profiling of microbiomes

Received: 10 September 2025

Accepted: 16 April 2026

Published online: 18 May 2026

Cite this article as: Visci G., Notario E., Defazio G. *et al.* Benchmarking short- and long-read sequencing technologies for metagenomic profiling of microbiomes. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-49725-3

Grazia Visci, Elisabetta Notario, Giuseppe Defazio, Mariano Francesco Caratozzolo, Sharon Natasha Cox, Bruno Fosso, Marinella Marzano & Graziano Pesole

Benchmarking Short- and Long-Read Sequencing Technologies for Metagenomic Profiling of Microbiomes

Grazia Visci[1†], Elisabetta Notario[2†], Giuseppe Defazio[1†], Mariano Francesco Caratozzolo[2], Sharon Natasha Cox [1], Bruno Fosso[1*], Marinella Marzano[2*], Graziano Pesole[1,2,3]

[1] Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, 70125 Bari, Italy.

[2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, 70126 Bari, Italy.

[3] Consorzio Interuniversitario Biotecnologie, 34148 Trieste, Italy.

[†] These authors contributed equally to this work.

*** Correspondence: bruno.fosso@uniba.it (BF); m.marzano@ibiom.cnr.it (MM)**

ABSTRACT

Background: Two culture-independent methods, amplicon-based sequencing and shotgun metagenomics, have significantly advanced the study of microbial communities. To date, short-read sequencing technologies have enabled high accuracy and deep coverage, while long-read sequencing approaches are increasingly being applied to improve genome assembly, despite challenges related to sequencing errors and nucleic acid input requirements. In this benchmark study, we compared the shotgun metagenomics approach across three sequencing technologies, Illumina (short reads), PacBio and Nanopore (long reads), using a 20-species commercial mock microbial community with even species representation. Specifically, we evaluated the effectiveness of the data generated by each platform in reconstructing genomes and identifying specific known taxa, as well as in understanding their functional potential, considering annotated genes, the length of predicted proteins and the number and types of inferred functions.

Results: Illumina sequencing provided high-throughput and high-quality data, but its limited read length precluded complete genome assembly. This affected the functional analysis, leading to an underestimation of coding and non-coding genes.

Nanopore sequencing yielded the longest reads, resulting in more contiguous assemblies, although it was affected by higher error rates and the choice of assembly method. PacBio offered the best balance between read length and base accuracy, but with a lower number of reads. This affected genome coverage for certain taxa, influencing the quality of their assemblies, the completeness of MAGs (Metagenome Assembled Genomes), and the accuracy of functional annotation. Nevertheless, PacBio successfully retrieved MAGs for all mock community species, and the genome annotation was consistent with the reference.

Conclusions: Evaluating the strengths and limitations of different NGS technologies and assembly strategies, this benchmark provides a practical framework for selecting the most suitable approach for optimizing data quality in microbiome genome characterization, according to study-specific goals.

Keywords: Shotgun metagenomics, microbiome, next-generation sequencing, third-generation sequencing, MAGs, functional analysis, mock community analysis.

1. INTRODUCTION

Exploring the taxonomic and functional biodiversity of microbial communities is essential for understanding ecosystem complexity, considering both the organisms and their roles. Microbial communities largely populate environmental or host-related niches and include bacteria, archaea, fungi, protists and viruses. As traditional approaches, relying on isolation in culture of microorganisms, principally prokaryotes, may uncover only about 1% of microbial biodiversity (1), DNA-sequencing based technologies have represented a revolutionary breakthrough. In the last two decades, high throughput sequencing technologies (HTS) have significantly enhanced our understanding of microbial communities and their essential roles in ecosystems as well as in human, animal, and plant health (2–4), paving the way to the so-called metagenomics approaches, such as amplicon-based (or DNA-metabarcoding) and shotgun metagenomics. Amplicon-based metagenomics relies on the selective amplification and sequencing of specific target genes (i.e., 16S or 18S rRNA genes, ITS) to obtain the taxonomic profile of microbial communities, although the choice of the target region can influence the results (2,5). Conversely, shotgun metagenomics involves the random sequencing of the entire genetic content of these communities providing not only taxonomic but also functional information (2,5). Both methods are

valuable for studying and characterizing microbiomes, each offering distinct advantages and being chosen based on the specific research question as well as cost considerations, with shotgun approach being considerably more expensive (3,5). However, while DNA barcodes range from 100 to 1,600 bp in length, a prokaryotic genome has an average size of around 5 Mbp, making shotgun metagenomics intuitively the most informative approach. Moreover, findings from shotgun metagenomics studies suggest that various microbiome related interactions, such as horizontal gene transfer, genetic content networks and microbiota-dependent metabolites, can have significant implications for the host-microbiome relationship (6–8). The ability to explore these interactions alongside taxonomic assignment, helps shed light on the human microbiome in both health and disease reveling molecular drivers of diseases, the spread of antibiotic resistance, disease-associated genetic elements, individual health and resilience (6–8). Equally important are the implications in environmental contexts, where metagenomic analysis helps explain why some ecosystems are more susceptible than others to disturbance or, conversely, more responsive to ecological restoration and sustainable management (9,10). Shotgun metagenomics enables high-resolution profiling of microbial communities, including taxonomic assignment at the strain level, the identification of unknown species through *de novo* assembly and the investigation of gene content, functional potential, and genomic plasticity (11–13). Nevertheless, genome assembly remains challenging because individual genomes must be reconstructed from a complex mixture of sequences derived from multiple organisms (14). Assembly-based metagenomic analysis can follow either a reference-based or an assembly-centric approach. The former relies on the alignment of sequencing reads to curated reference databases to directly quantify microbial taxa and functions (15,16) and it is particularly powerful for comparing community composition and functional potential across multiple samples or conditions. The latter approach is based on the direct assembly and binning of sequencing reads to reconstruct metagenome-assembled genomes (MAGs). The reconstruction of MAGs represents a critical step in assembly-based metagenomics and is computationally intensive (17). Indeed, genomes from different bacteria may share highly similar regions, and only when they are fully assembled into gapless, circularized genomes do they become comparable to genomes obtained from isolated and pure cultures, thereby enabling classification up to the strain level (18). Moreover, an additional challenge arises when dealing with uncultured and uncultivable organisms that

have not yet been sequenced and are therefore absent from reference genome databases. This leads to an increased proportion of reads that cannot be aligned and are consequently classified as unassigned (14). Nevertheless, the assembly-centric approach remains especially valuable for the discovery of novel taxa and uncharacterized functions, with outcomes that may be strongly influenced by the choice of sequencing strategies and platforms. In this context, shotgun metagenomics may face technical limitations, which can affect assembly quality and genome reconstruction. Low DNA concentrations in metagenomic samples may lead to the use of amplification protocols, increasing experimental bias (19,2). Additionally, host-DNA interference can reduce the sensitivity of detecting low-abundance bacterial species. To mitigate this issue, higher sequencing depth is required, which in turn increases overall sequencing costs needed to achieve adequate microbial genome coverage (20–22). Indeed, a more in-depth characterization of microbial communities requires HTS platforms, that include short- and long-read sequencing technologies, able to produce large amounts of data. Short-read sequencing has dominated microbiome studies until now, thanks to its high-quality reads, low-input protocols and high coverage (23–27), despite requiring fragmentation and amplification steps (28). On the other hand, long–read sequencing technologies can yield long and ultra-long reads directly from single DNA molecules, despite the lower per-base accuracy and the higher amount of DNA input required (21,23,29–31). Several benchmark studies have been conducted over time, comparing second- and third-generation sequencing platforms (23,30–32). These studies serve as essential resources for researchers to better understand the advantages and limitations of each technology in reference- and assembly-based shotgun metagenomics (31).

In this study, we applied an assembly-based shotgun metagenomics approach to a microbial community with known composition (mock community). We adopted Illumina NovaSeq 6000 for short-read sequencing, alongside the PacBio Sequel System IIe and Oxford Nanopore GridION for long-read sequencing. Using a commercially available prokaryotic mock community, we established a controlled benchmarking framework to assess sequencing accuracy, coverage and assembly efficiency. Beyond evaluating the limitations of individual sequencing protocols and the ability of different assemblers to reconstruct high-quality MAGs, a key added value of this study is the investigation of microbial taxonomic assignment and gene annotation derived from the recovered genomes. Here, we provide a test case using standards characterized by intra- and inter-species diversity to

demonstrate how current technologies can be applied to explore multiple aspects of the microbial “dark matter”.

2. MATERIALS AND METHODS

2.1. Mock community sample

The commercial mock microbial community, ATCC 20 Strain Even Mix Genomic Material (MSA-1002, ATCC, USA, https://www.atcc.org/products/msa-1002), was used as a benchmark for the shotgun metagenomic study. It consists of a mix of genomic DNA derived from 20 fully sequenced, characterized, and authenticated ATCC Genuine Cultures (5% for each strain) (**Supplementary Table S1**). Fluorometric quantification and genome quality were assessed using the Qubit dsDNA HS assay (Thermo Fisher Scientific, Waltham, MA, USA) and the Genomic DNA 165 kb Kit for the Femto Pulse System (Agilent, Santa Clara, CA, USA) (**Supplementary Figure S1A**), respectively. The total yield of the commercial purchased sample was approximately 200 ng. Therefore, three genomic DNA mixes from the same production batch (Lot. 70001383) were used for the different applications.

2.2. WGS library preparation and sequencing

Different shotgun metagenomic protocols and sequencing platforms were used in this study. The mock community DNA was used as input for library preparation and sequenced on NovaSeq 6000 (Illumina, San Diego, CA, USA), GridION (Oxford Nanopore Technologies, Oxford, UK) and the Sequel IIe System (PacBio, Menlo Park, CA, USA). We used the same input DNA amount as a normalization factor for cross-platform comparison. Moreover, one sequencing unit was assigned per platform, one flow cell for Illumina and ONT, and one SMRT Cell for PacBio. The mock sample was either multiplexed with other samples or sequenced individually to maximize sequencing capacity. For each platform, the sequencing output and the number of samples multiplexed per flow cell or SMRT cell are reported below.

Illumina Sequencing

Illumina DNA Prep kit was used, starting from 200 ng of DNA of the mock community, following the protocol instructions (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/illumina-dna-prep-reference-guide-1000000025416-09.pdf). The protocol uses bead-

linked transposases to tagment DNA, generating an insert size of ~350 bp, and then includes a step of amplification of the tagmented DNA. All the libraries were quality checked through High Sensitivity DNA Assay for 2100 Agilent Bioanalyzer (Agilent, Santa Clara, CA, USA) and quantified using the Qubit dsDNA HS Assay (Thermo Fisher Scientific, Waltham, MA, USA). The library was sequenced on the NovaSeq 6000 with a 2 × 150 bp paired-end sequencing layout (NovaSeq 6000 S4 Reagent Kit v1.5 - 300 cycles). The mock sample was loaded multiplexed with 60 other samples to maximize the sequencing capacity of the single S4 flow cell (maximum flow cell output 3Tb). Approximately 16.3 Gb of data were generated from the mock sample.

Nanopore Sequencing

Approximately 200 ng of DNA were used as input for the Genomic DNA Ligation Sequencing Kit (ONT SQK-LSK114) (https://nanoporetech.com/document/genomic-dna-by-ligation-sqk-lsk114?device=GridION) and sequenced on the GridION platform, without fragmentation or amplification steps. The sample was loaded individually on a single MinION Flow Cell (R10.4.1, maximum flow cell output, 50Gb). A total of 2.8 Gb of data was generated for the mock sample.

PacBio Sequencing

Library preparation was performed following the PacBio procedure and checklist: “Preparing whole genome and metagenome libraries using SMRTbell Prep Kit 3.0” (PN 102-166-600 - APR2022) starting from about 200 ng of fragmented DNA. According to the manufacturer's instructions, the DNA of the mock community was sheared at speed 35 by the Megaruptor3 (Hologic, Inc). The genomic profile of fragmented DNA was assessed with the Genomic DNA 165 kb Kit for the Femto Pulse System (Agilent, Santa Clara, CA, USA) (**Supplementary Figure S1B**). Then, the Binding Kit 2.2, Internal Control 1.0 and Sequel II Sequencing Kit 2.0 were used for sequencing on the PacBio Sequel IIe System. The mock sample was sequenced together with 5 other multiplexed samples on a single SMRT Cell 8M (maximum SMRT Cell output 30 Gb). PacBio sequencing produced approximately 0.8 Gb of data for the mock sample.

2.3. Raw data trimming, assembly and, mapping on reference genomes

Illumina data analysis and assembly

Illumina raw sequencing data were initially quality checked using FastQC (v0.11.9) and low-quality reads were trimmed by using trimmomatic (v0.39, PE ILLUMINACLIP LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50) (33). Trimmed data were assembled using two alternative approaches: MEGAHIT (v1.2.9, --k-list 21, 29, 39, 59, 79, 99, 119, 141 --k-step 10 --min_count 2) (34) and metaSPAdes (v3.15.5, --meta -k 21,29,39,59,79,99,119 -m 500 --phred-offset 33) (35).

Nanopore data analysis and assembly

Raw Nanopore sequencing data were initially quality checked using pycoQC (36). Porechop_ABI (v0.5.0, --ab_initio --format fastq.gz) (37) was used to identify and trim adapter sequences. Trimmed data were assembled using metaFlye (v 2.9.2-b1786, --nano-raw –meta -i 5) (38) and metaMDBG (v1.0, asm –in-ont) (39).

PacBio data analysis and assembly

PacBio HiFi data were initially quality checked using FastQC (v0.11.9). Then, cutadapt (v4.5, --overlap 35 -e 0.1 --discard -j 5 --revcomp) (40) was applied to check the HiFi reads for adapter presence. HiFi reads containing adapters were discarded and excluded from subsequent analysis. Trimmed data were assembled using metaFlye (v2.9.2-b1786, --pacbio-hifi --meta -i 5) (38) and metaMDBG (v1.0, asm –in-hifi) (39).

2.4 Mapping to reference genomes and reference coverage

Sequencing data were mapped to the 20 prokaryotic strain reference genomes downloaded from ATCC (https://www.atcc.org/, accessed on 1 March 2021) using minimap2 (v2.26-r1175). The following presets were applied: Illumina (-ax sr), Nanopore (-ax map-ont -L), and PacBio (-ax map-hifi -L). Through samtools (v1.3.1), SAM files were into BAM format and sorted. Finally, sorted bam files were used to measure genome coverage and sequencing depth through the samtools coverage function (-ff 1284, to exclude unmapped reads and secondary alignments, -d 0, to avoid any limits in coverage counts).

2.5 Assembly evaluation, binning and bin refinement

The obtained assemblies were evaluated using metaQUAST (v5.2.0, default parameters) (41) with the -r option to map contigs to reference genomes. Seqkit (v2.8.2, stats -j10 -t -a) (42)was used to retrieve overall statistics for obtained contigs.

Regardless of the sequencing and assembly approach, the obtained contigs were binned and the obtained bins were refined using metaWRAP (43). Initial binning was performed using metaBAT2 (v2.12.1, min contig length 1500) (44), MaxBin2 (2.2.4, min contig length 1000) (45) and CONCOCT (v1.0.0, min contig length 1000) (46). During the bin refinement process, inferred MAGs were quality checked using CheckM (v1.0.18) (47) and genomes with completeness ≥ 90% and contamination ≤ 5% were marked as high quality, those with completeness ≥ 50% and contamination ≤ 10% as medium quality, and all others as low quality (17).

2.6 MAGs comparison to reference genomes

The obtained MAGs were compared to the ATCC reference genome using MASH (v2.3, sketch -k 21 -s 15000) (48). Both reference genomes and MAGs were sketched using 15,000 minhashes and “all versus all” comparisons were performed. Moreover, a phylogenetic comparison of the MAGs was carried out using GTDB-tk (v2.1.1) and the GTDB reference database (r214) (49). Finally, MAGs were taxonomically classified by using kMetaShot (v2.0, default options) (18).

2.7 MAGs dereplication

The obtained MAGs with at least medium overall quality and the ATCC reference genomes were dereplicated using the dRep (v3.5.0, dereplicate --ignoreGenomeQuality --genomeInfo) (50). Considering the presence of two pairs of co-generic species in the employed mock, the Ward algorithm (51) for hierarchical clustering was applied, to minimize within-cluster variance.

2.8 MAGs gene annotation

Annotation of both inferred MAGs with at least medium overall quality and reference genomes downloaded from ATCC was performed using Bakta (v1.4.0, --min-contig-length 200) (52). The protein length profile inferred in MAGs was compared to that of ATCC reference genomes by performing pairwise Wilcoxon tests. Proteins labelled as hypothetical were excluded from these comparisons. Annotated protein products were compared between reference genomes and obtained MAGs both qualitatively, by counting common predicted protein functions (i.e., considering the product description assigned by Bakta to the CDS) and those private to reference genomes and MAGs, and quantitatively by measuring the Jaccard distance. Jaccard distance was measured using an *in-house* developed Python script.

Predicted protein lengths were qualitatively compared between MAGs and reference genomes using boxplots and quantitatively through Wilcoxon test. Furthermore, the effect size of the variation was measured as the change between the relative difference in mean protein lengths between MAGs and the reference genomes ($d_{\%}$). It represents a normalized measure of difference in observed protein lengths. It was measured as follows:

$$d_{\%} = \frac{\overline{L}_{MAG} - \overline{L}_{REF}}{\overline{L}_{REF}} \times 100$$

ATCC reference genome ribosomal RNA sequences were extracted form Bakta annotation and searched with BLAST (2.16.0, build Mar 28, 2025, -query Ref_rRNA_seq.fa -out rRNA_blast_search.out -subject mags_rrna_seq.fa -word_size 11 -outfmt 6 -sorthits 3 -perc_identity 97 -qcov_hsp_perc 90). Only matches achieving at least 97% sequence identity and 90% query coverage were retained.

2.9 MAGs quantification

Considering the variability in genome size among the species in the mock community (**Supplementary Table S1**) and the fact that an equal amount of genomic DNA was added to the mix, the number of expected genome copies was estimated to infer the expected relative abundances. We estimated the mass of each genome in ng (nGM_i), considering that the average weight of a base pair in dsDNA is 607.4 g/mol.

Finally, we estimated the Genome Copy Number for each species *i* (GCN_i) considering that the same amount of genomic DNA was added to the mixture (i.e. 10 ng):

$$GCH_i = \frac{10}{nGM_i}$$

The estimated genomic copies for each species and the corresponding relative abundances are shown in **Supplementary Table S1**.

Subsequently, trimmed reads were mapped to the obtained MAGs using minimap2 (using the same options listed in section 2.4). MAGs coverage was estimated by using the samtools coverage function (-ff 1284, to exclude unmapped reads and secondary alignments, -d 0, to avoid any limits in coverage counts).

3. RESULTS

3. 1 Sequencing throughput

The sequencing data obtained for each technology (Illumina, Nanopore and PacBio), before and after adapter trimming, are shown in **Table 1**.

***Table 1**: **Statistics regarding raw and trimmed sequencing data.** For each sequencing technology, the following data are shown: i) N. of seqs: number of produced reads; ii) Yield: total throughput in bases; iii) Min len: minimum sequence length in bp; iv) Avg len: average read length in bp; v) Median len: median read length; vi) Max len.: maximum sequence length in bp; vii) N50: half of the sequenced bases are contained in sequences of the this length or longer ; viii) Q20 (%) amount of called basis with quality score ≥ 20; ix) Q30 (%) amount of called basis with quality score ≥ 30.*

Seq technologies	N. of reads	Yield	Min len	Avg len	Median len	Max len	N50	Q20 (%)	Q30 (%)
Illumina									
raw data	115,746,688	16,353,634,835	35	141.3	151	151	151	94.01	85.70
trimmed data	108,028,131	15,066,924,022	50	139.5	151	151	151	95.54	87.84
Nanopore									
raw data	2,301,340	2,806,148,731	5	1,219.4	489	918,116	3,051	61.56	48.43
trimmed data	2,299,453	2,724,939,449	1	1,185	455	918,116	3,194	62.39	49.30
PacBio									
raw data	111,629	805,182,430	258	7,213	6737	24,210	9,001	99.24	98.37
trimmed data	111,626	805,164,864	258	7,213.1	6737	24,210	9,001	99.24	98.37

Both Nanopore and PacBio produced reads with an average length of approximately 1 kb and 7 kb, respectively. Nanopore produced the longest read, spanning about 0.9 Mb. Considering the amount of retained sequences/bases, 93.33%/92.13% passed the trimming step for Illumina sequencing, while the corresponding values were 99.91%/97.10% for Nanopore and 99.99%/99.99% for PacBio. Considering the average quality of the called bases, the higher Q20 and Q30 values (i.e. the number of bases with a quality score ≥ 20 and ≥ 30, respectively), were observed in PacBio data, followed by Illumina, while Nanopore sequences obtained the lowest ones. Furthermore, trimming improved the Q20 and Q30 values with the only exception of PacBio data.

3.2 Sequencing depth and genome coverage

Before performing the assembly, trimmed reads were mapped to reference genomes to evaluate the average coverage and sequencing depth. Considering M as the length of a selected reference genome and N as the number of genomic bases covered at least one time by sequencing reads, genome coverage is defined

as N/M × 100. Coverage depth, or sequencing depth, describes the number of times a base in a reference genome or assembly is covered by sequencing reads. Because this number can vary along the reference genome, descriptive parameters of its distribution, or the distribution itself, can be used to characterize coverage depth.

Initially, we estimated the mean coverage across reference genomes (**Figure 1**). The three sequencing technologies produced variable sequencing depth, with Illumina yielding the highest values (median 441.01, IQR = 152.17, mean 1280.96), approximately two orders of magnitude higher than Nanopore (median 29.91, IQR = 106.71, mean 88.56) and PacBio (median 12.84, IQR = 5.7, mean 18.43). Moreover, all the mock genomes were completely covered by Illumina (median 100, IQR = 0, mean 94.43) and Nanopore (median 100, IQR = 0, mean 94.43) (**Figure 1**). PacBio (median 100, IQR = 0, mean 93.43) fully covered 19 out of 20 genomes (**Figure 1**). Specifically, *Schaalia odontolytica* showed an overall coverage of approximately 92.4% with an average depth of 2.7X using PacBio, compared to 384X and 28X for Illumina and Nanopore sequencing, respectively (**Supplementary Table S2**).

Figure 1: Distributions of sequencing depths for reference genome contigs with length of at least 100 bp obtained with Illumina, Nanopore and PacBio sequencing platforms. Each dot represents a contig and is colored according to its coverage.

Moreover, we also evaluated the coverage depth (or sequencing depth) by measuring the proportion of reference genomes covered at 20X, 30X, 40X and 50X (**Figure 2**). Using Illumina sequencing, all the genomes were completely covered at 50X. When considering long-read sequencing, only *Escherichia coli* sequenced with Nanopore was completely covered at 40X. Overall, for 4 species (namely *Streptococcus agalactiae*, *Streptococcus mutans*, *Phocaeicola vultgatus* and *Deinococcus radiodurans*) we observed that less than 50% of the genomes were covered at 20X. Finally, with PacBio sequencing, none of the genomes was completely covered at 20X.

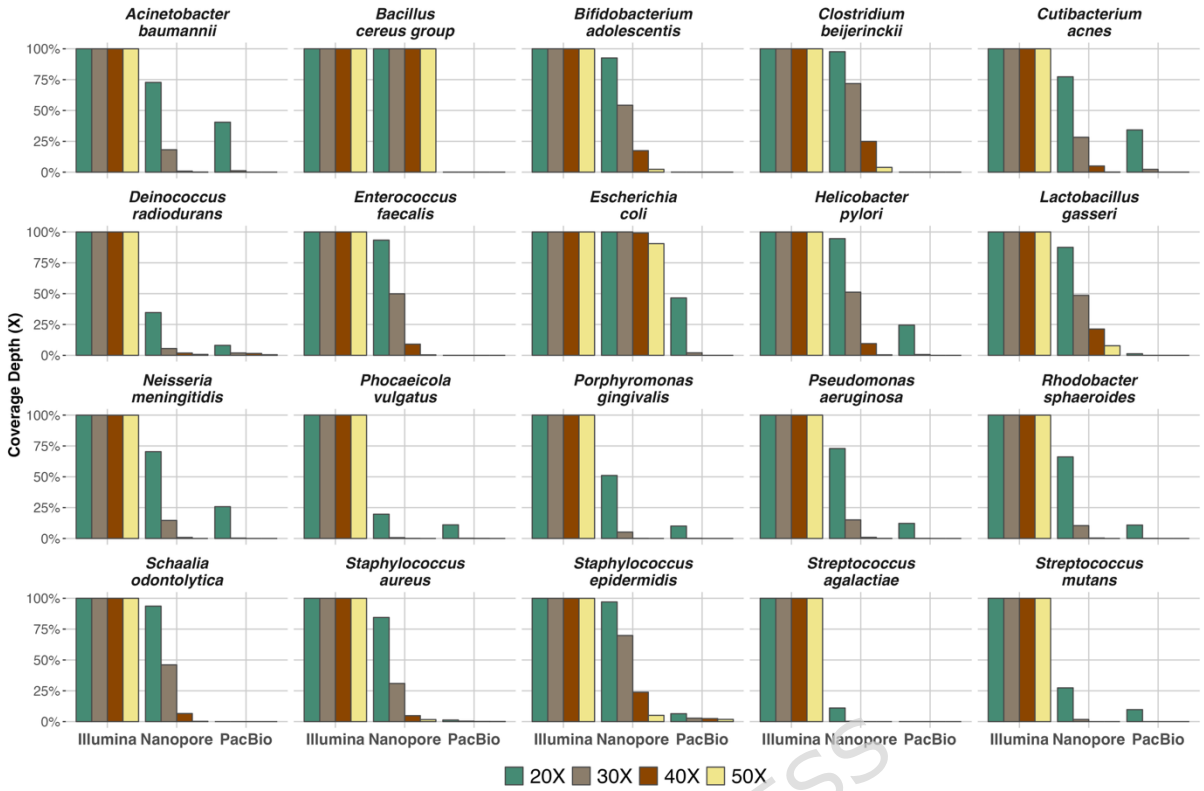

Figure 2: For each reference genome, coverage depth was calculated as the proportion of the bases of the genome covered at a minimum depth of 20x, 30x, 40x, and 50x. *The barplots represent the relative amount of the genome covered at the specific sequencing depth.*

3.3 Assembly Evaluation

Two different assembly algorithms were used for each sequencing technology: MEGAHIT and metaSPAdes for short-reads, and metaFlye and metaMDBG for long-reads. The assembly summary statistics are shown in **Table 2**.

Table 2: Assembly statistics for MEGAHIT, metaSPAdes, metaFlye and metaMDB: *i) Contig: number of produced contigs; ii)Tot Len: assembly total length in bp; iii) % Expected Tot Len: obtained fraction (%) of the sum of lengths for reference mock genomes; iv) Min length: shortest contig length in bp; v) Avg len: average contigs length in bp; vi) Max len: maximum contig length in bp; vii) N50; viii) GC(%).*

Assembly	Contig	Tot Len	% Expected Tot Len	Min len	Avg len	Max len	N50	%GC
Illumina								
MEGAHIT	1,135	65,990,405	98.43%	773	58,141	955,220	174,256	47.10

metaSPAdes	2,144	66,587,045	99.32%	120	31,057	1,405,032	232,008	47.10
Nanopore								
metaFlye	113	67,286,575	100.37%	1,002	595,456.40	6,374,455	2,227,176	47.18
metaMDBG	4,838	80,148,321	119.55%	267	16,566.40	4,642,452	41,003	47.19
PacBio								
metaFlye	358	60,826,078	90.73%	3,125	169,905.20	6,374,538	1,841,921	47.09
metaMDBG	581	65,931,520	98.34%	1,097	113,479.40	6,374,527	2,032,857	47.06

Considering assembly contiguity, Illumina data obtained N50 in the range of hundreds kilobases with both assemblers. The widest contig length was obtained with metaSPAdes, ranging from 120 bp to 1.4 Mbp (**Table 2**). Regarding long-reads, the same assembly algorithms behaved differently depending on the analysed data. When using Nanopore data, metaFlye achieved a markedly higher contiguity than metaMDBG, with an N50 of about 2.2 Mbp compared with only 41 Kbp for metaMDBG. Moreover, metaMDBG produced the largest number of contigs, including very short ones (267 bp) indicating a much more fragmented assembly (**Table 2**). However, for PacBio data the N50 values obtained with metaFlye and metaMDBG were similar, both reaching at least 2 Mbp, although metaMDBG produced a moderately higher number of contigs (581 vs 358). In this case, both assemblers were able to produce contigs longer than 6 Mbp (**Table 2**). The obtained contigs were evaluated using MetaQUAST, comparing metagenome assemblies based on alignments to the closest reference genome. In this section, we use the term coverage to refer to the proportion of the reference genomes that is covered by the obtained contigs. The number of contigs mapping to each reference genome and the observed coverage are shown in **Figure 3**. Overall, regardless of the assembly algorithm used, the contigs obtained from assembling Illumina reads broadly map to the reference genomes, covering them almost completely. Nonetheless, a higher number of contigs mapping to the reference genome was generally observed in these assemblies compared to those generated from long reads (**Figure 3**). The lowest reference genome coverage (92.6%) was observed for *Porphyromonas gingivalis* using MEGAHIT, which produced 104 contigs. In contrast, *Pseudomonas aeruginosa* achieved the highest coverage (99.6%) with just 27 contigs using metaSPAdes (**Supplementary Table S3**).

Notably, *Cutibacterium acnes* was assembled with 99.4% coverage and the lowest number of contigs (10) using both assemblers.

For Nanopore data, metaFlye produced the most contiguous assemblies with 4 out 20 genomes assembled at 100% coverage of the reference genome. Specifically, the genome of *Escherichia coli* was assembled into a single contig mapping completely to the corresponding ATCC reference genome. Among the remaining genomes, 7 were assembled with > 99.9% coverage, including 3 single-contig assemblies, and 9 were assembled with > 98.2% coverage, each comprising at least 3 contigs. By contrast, metaMDBG produced the most fragmented assemblies from Nanopore data (**Supplementary Table S3**).

Considering PacBio sequencing data, 4 out of 20 genomes were assembled at 100% by metaFlye, including the genome of *E. coli* with only 1 contig. In this case, only 3 genomes were reconstructed at ≥ 99.9%, whereas 10 genomes were assembled at > 90%. *Bifidobacterium adolescentis*, *Bacillus pacificus* and *Schaalia odontolytica* had the least complete assemblies, covering 60.1%, 48.9% and 14.1% of the reference ATCC genomes, respectively. MetaMDBG produced different results with PacBio data: 6 out 20 ATCC reference genomes were covered at 100%, including *Cutibacterium acnes, Escherichia coli, Helicobacter pylori, Pseudomonas aeruginosa* and *Streptococcus mutans,* each assembled in a single contig. Among the remaining genomes, 4 were assembled at >99.9%, and 8 at > 95.6% with at least 3 contigs. Finally, compared to the metaFlye assemblies, the percentage of assembled genome coverage of *B. pacificus* and *S. odontolytica* increased to 92.5% and *76.7%*, respectively (**Supplementary Table S3)**. The largest contig obtained for both long-read sequencing approaches, Nanopore and PacBio, corresponded exactly to the *Pseudomonas aeruginosa* genome (6,374,461 bp).

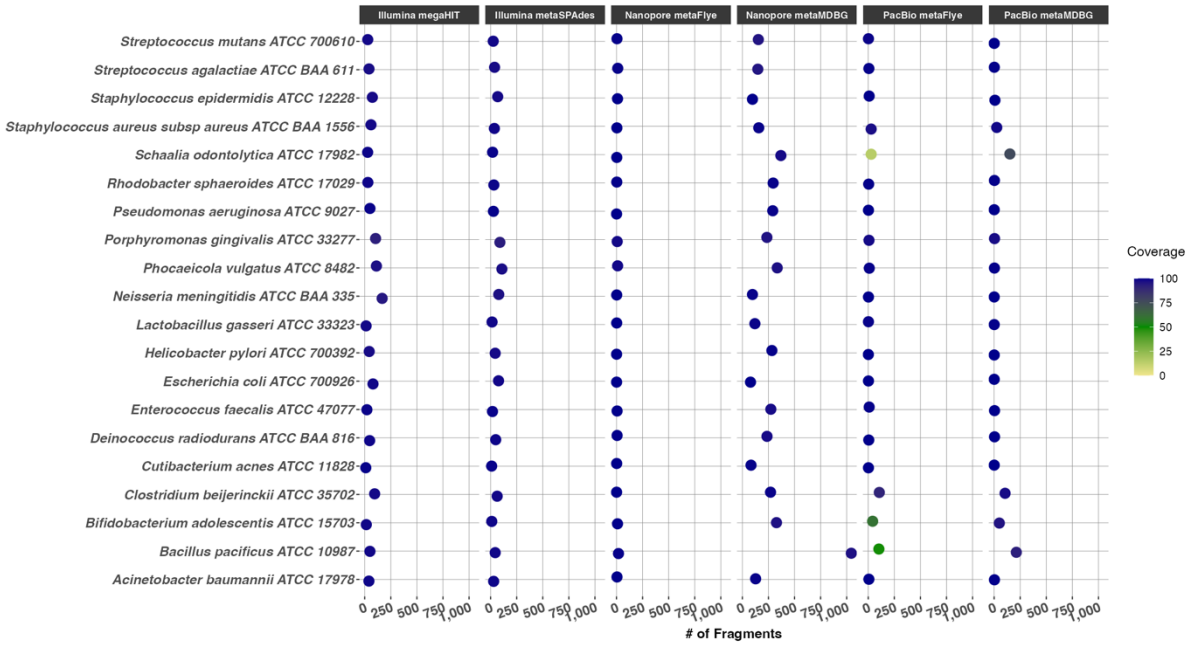

Figure 3: Observed number of fragments (x axis) mapping on the reference genomes (y axis) and observed genomes coverage for combinations of sequencing platforms and assemblers (panels).

3.4 MAGs evaluation

Subsequently, we performed contig binning to generate MAGs, a procedure that groups assembled contigs into bins representing putative genomes, typically corresponding to microbial species or strains (53). It is a critical step that enables the reconstruction of draft genomes from complex microbial communities, allowing further analysis, such as the investigation of the metabolic potential and taxonomic classification (52).

Contigs binning and refinement were performed using MetaWRAP which integrates bins generated by MetaBAT2, CONCOCT and MaxBin2 to produce refined MAGs. Then MAG quality was evaluated using CheckM and genomes with a completeness ≥ 90% and contamination ≤ 5% were marked as high quality; those with completeness ≥ 50% and contamination ≤ 10% as medium quality; and all others as low quality. The results obtained are summarized in **Table 3**. MAGs were taxonomically annotated using kMetaShot and a phylogenetic tree including reference ATCC genomes was built through GTDBtk.

Table 3: Summary of the number and quality of obtained MAGs for each assembly approach.

High quality Genomes: completeness ≥ 90% and contamination ≤ 5%; Medium quality Genomes: completeness ≥ 50% and contamination ≤ 10%; Low quality Genomes: all remaining MAGs not meeting the criteria for high or medium quality.

Assembly	***Number of MAGs***	***Quality of MAGs***
Illumina		
MEGAHIT	18	16 high; 2 medium
metaSPAdes	20	15 high; 5 medium
Nanopore		
metaFlye	18	18 high
metaMDBG	24	4 high; 13 medium; 7 Low
PacBio		
metaFlye	19	17 high; 1 medium; 1 Low
metaMDBG	20	19 high, 1 medium

kMetaShot was able to classify all the obtained MAGs, regardless of their quality, and the obtained classification corresponded to the expected species. Nonetheless, it is worthy to note that *Rhodobacter sphaereoides* and *Propionibacterium acnes* were renamed as *Cereibacter sphaeroides* (55) and *Cutibacterium acnes* (56), respectively. Finally, *Bacillus pacificus ATCC 10987* is annotated in the NCBI taxonomy as *Bacillus cereus ATCC 10987*, was concordantly labeled by kMetaShot.

All 20 expected genomes were correctly recovered only when assembling Illumina data with metaSPAdes and PacBio data with metaMDBG (**Table 3**). In contrast, when using MEGAHIT we were unable to retrieve the genomes of *Streptococcus agalactiae* and *Staphylococcus epidermidis*, both belonging to genera represented by two species in the mock community (**Figure 4)**. With Nanopore data processed by metaFlye we obtained 18 high quality MAGs, but the two *Staphylococcus spp.* were missing. All the expected species were retrieved from the metaMDBG assemblies, however according to kMetaShot classification two MAGs were observed for *Cutibacterium acnes* (high and low), *Helicobacter pylori* (medium, low), *Porphyromonas gengivalis* (both low) and *Staphylococcus epidermidis* (medium, low). This resulted in an overestimation of the number of MAGs, increasing it from 20 to 24 (**Table 3, Figure 4**). Regarding PacBio data, using metaFlye the only missing species was *Schaalia odontolytica.* When assembling PacBio data with metaMDBG, all the expected genomes were retrieved, 19 high-

quality MAGs and 1 medium-quality MAG, represented by *Schaalia odontolytica* (**Table 3, Figure 4**).

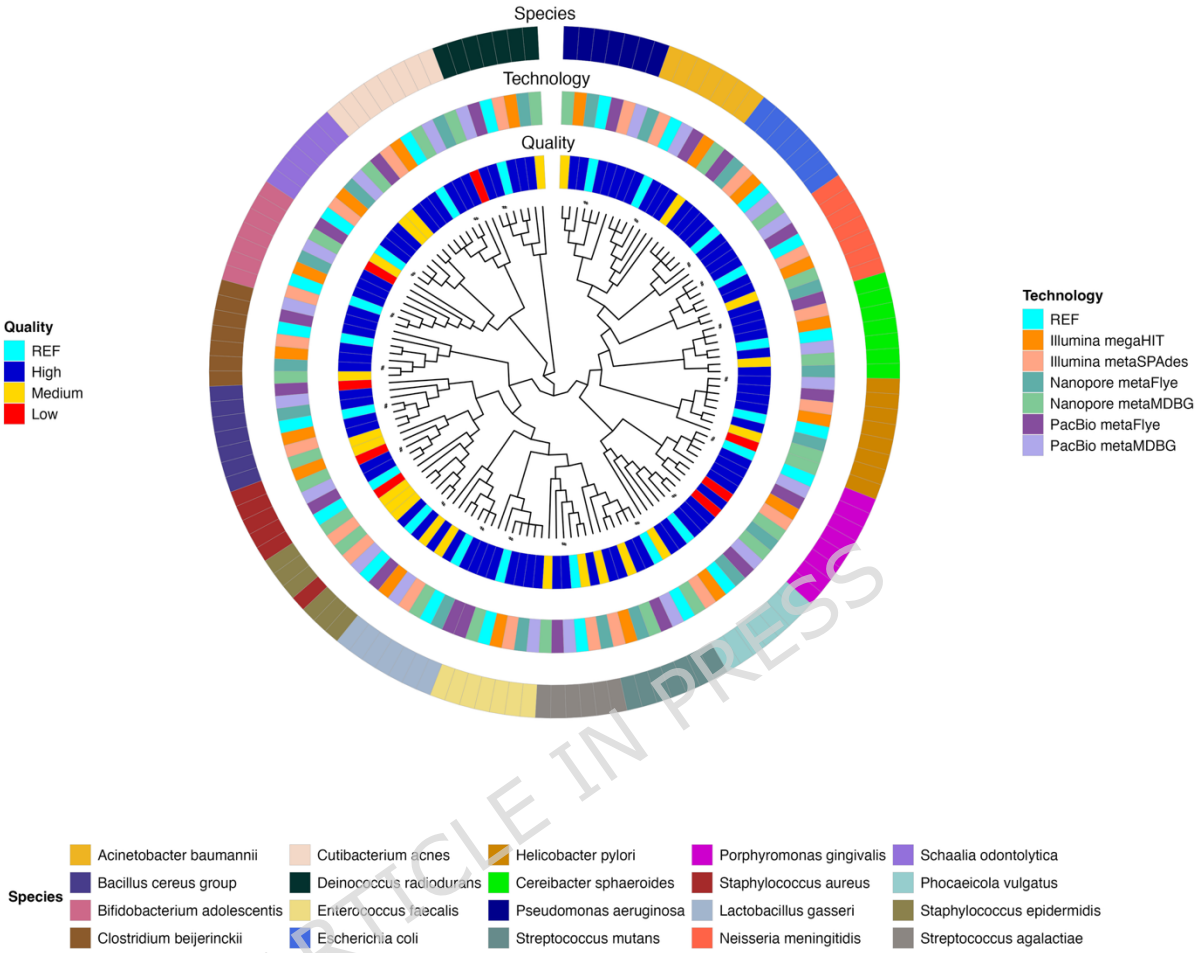

Figure 4: Tree representing the phylogenetic relationship among the observed Hybrid MAGs and ATCC reference genomes. *Circular heatmaps correspond to MAG quality (Quality), the combination of sequencing approach and assembly method (Technology) and species scientific name (Species). dRep representative genomes are marked on tree leaves with *.*

Furthermore, considering that kMetaShot classification and GTDBtk phylogeny are supervised approaches relying on reference genomes collection and taxonomic information, we also performed unsupervised clustering and dereplication through ANI (Average Nucleotide Identity) using dRep. It relies on a two-step approach, applying a concise ANI inference through sketching with MASH to infer primary clusters (ANI ≥ 90%) followed by secondary clustering based on precise ANI estimation by fastANI (ANI ≥ 95%). Finally, following secondary clustering completion, the best genome for each cluster was selected by considering genomic features (completeness, contamination, size, and strain heterogeneity),

assembly and clustering quality metrics (N50 and centrality). Dereplication results excluding low-quality MAGs are shown in **Supplementary Table S4** and secondary cluster representative genomes are labelled in **Figure 4**. dRep was applied by pooling together the 20 ATCC reference genome with high- and medium- quality MAG sequences and identified 20 clusters, one for each mock species, and a complete correspondence between clusters and kMetaShot taxonomic classification was observed. Regarding the choice of the best genome per cluster, 7 out of 20 were chosen among the ATCC reference ones, while the other 13 were chosen among the MAGs (2 Nanopore + metaFlye, 6 PacBio + metaMDBG, and 5 PacBio + metaFlye).

The obtained MAG genome sizes were compared with those of the reference genomes (**Supplementary Figure S2**) using the kMetaShot classification as a guide. Regardless of sequencing technology and assembly methods, the genome sizes of *Rhodobacter sphaeroides* and *Deinococcus radiodurans* were overestimated compared with their reference genomes.

For short reads, the genome size of *Bifidobacterium adolescentis* was overestimated with both metaSPAdes and MEGAHIT, and *Bacillus pacificus* displayed the same trend when using MEGAHIT. The assembled genome sizes of *Bacillus pacificus*, *Bifidobacterium adolescentis*, *Helicobacter pylori*, and *Porphyromonas gingivalis* in metaFlye assemblies exceeded the expected size. By contrast, *Acinetobacter baumannii*, *Phoaecicola vulgatus, Staphylococcus aureus* subsp. *aureus* and *Staphylococcus epidermidis* MAGs showed a similar trend in PacBio assemblies regardless of the applied assembler (**Supplementary Table S5**).

Short-read sequencing allowed complete reconstruction with a 100% match to the reference of only 2 MAGs, whereas long-read sequencing approaches recovery of 8 complete MAGs from Nanopore data (metaFlye) and 8 complete MAGs from PacBio data (metaMDBG) (**Supplementary Table S5**).

Finally, considering the mock bacterial species belonging to the same genus, we observed that none of the MAGs obtained by assembling Illumina data using MEGAHIT were identified as *Staphylococcus epidermidis* or *Staphylococcus aureus*. In contrast, *Streptococcus agalactiae* and *Streptococcus mutans* were taxonomically identified, although only the latter had a genome size matching 100% of the reference genome. A different result was obtained when using Illumina data assembled with metaSPAdes. In this case, we identified both *Staphylococcus* and *Streptococcus* species, but with a lower match to their

respective reference genomes (*S. epidermidis* 55.2% and *S. aureus* 40.4%, *S. agalactiae* 82.7% and *S. mutans* 58.7%) (**Supplementary Table S5**). Regarding long-read sequencing approaches, the results obtained with Nanopore data were similar to those observed with short reads. Specifically, no MAGs derived from Nanopore data assembled with metaFlye were identified as *S. epidermidis* or *S. aureus.* Conversely, the MAGs classified as *S. agalactiae* and *S. mutans* matched the reference genomes at 99.3% and 98.8%, respectively. However, the analysis of Nanopore sequencing data using metaMDBG allowed the reconstruction of MAGs and classification of both *Staphylococcus* and *Streptococcus* species despite a lower correspondence to their reference genomes (**Supplementary Table S5**). In the case of PacBio data assembled with both metaFlye and metaMDBG, *Staphylococcus epidermidis* and *Staphylococcus aureus* MAGs were taxonomically classified, although with an overestimation in genome size compared to the reference. Similarly, *Streptococcus agalactiae* MAGs matched the reference genome at 99.7% (metaFlye) and 99.8% (metaMDBG), whereas *Streptococcus mutans* matched 100% (**Supplementary Table S5**).

Furthermore, we used MASH to measure the distance between the obtained MAGs and their corresponding reference genomes (**Figure 5**). Regardless of the sequencing technology or assembly approach applied, high-quality MAGs showed a distance from the reference genomes below 1%. For medium-quality MAGs, MASH distances within 2% were observed, with the only exception of *Staphylococcus aureus*. In this case, the two MAGs assembled from Illumina data (i.e., MASH distance of 1.9% with MEGAHIT and 8.7% with metaSPAdes) were the least similar to the reference genome, even when compared to the low-quality MAG assembled from Nanopore reads using metaMDBG (1.4%).

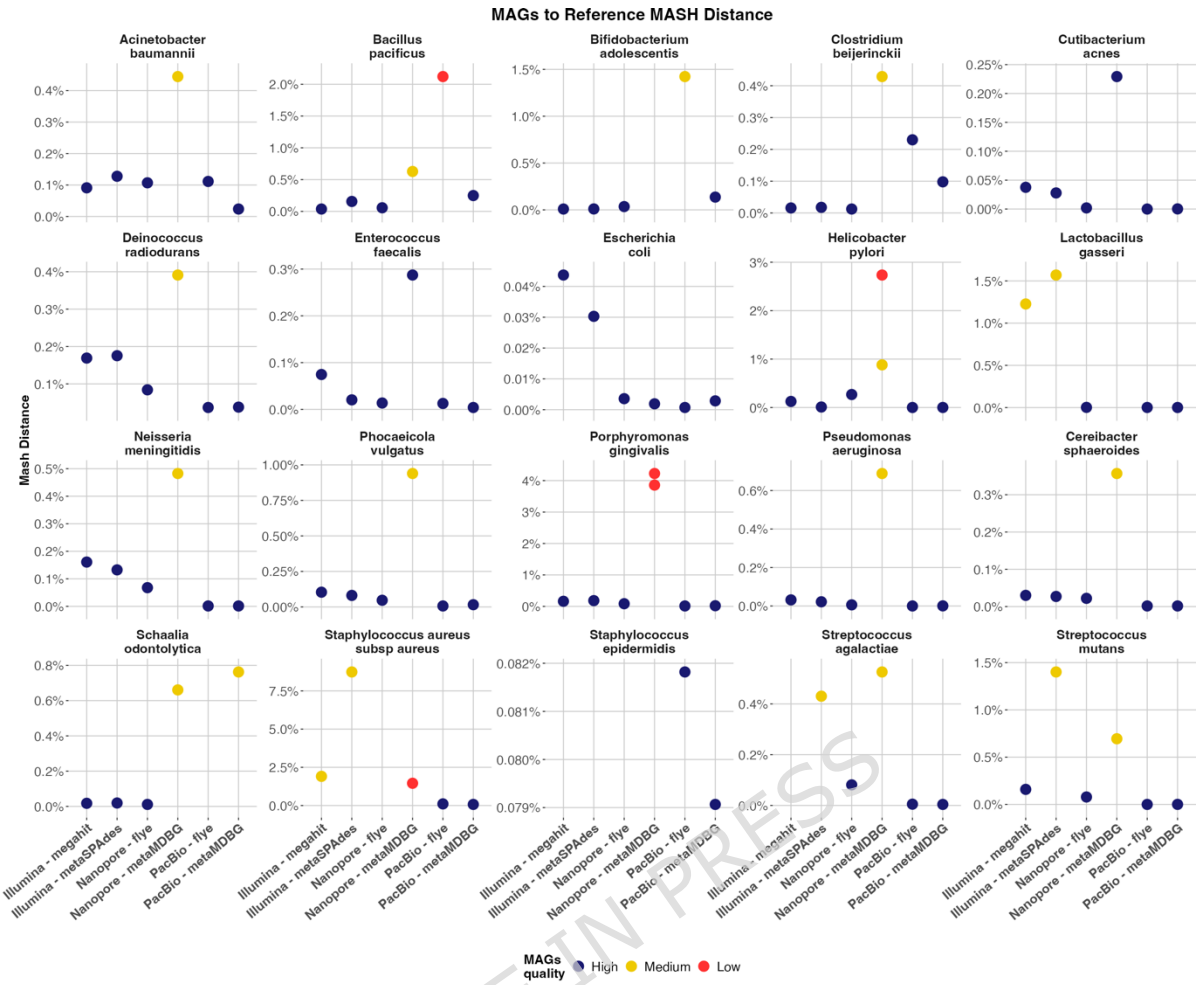

Figure 5: MASH distances between obtained MAGs and corresponding reference genome. *MASH-based 15,000 sketches of MAGs and corresponding reference genome were compared to infer an approximation of the ANI (Average Nucleotide Identity) distance. For each reference genome, the observed distances to the obtained MAGs were shown as a scatterplot. The combination of sequencing technique and assembly methods and the observed MASH distance as percentage were shown on the x and y axis, respectively. The dots are colored according to the quality of the MAGs.*

Finally, we compared the expected species abundances (**Supplementary Table S1**) with those inferred from MAG coverage and relative abundance estimates (**Supplementary Figure S3**) across each combination of sequencing technologies and assembly approaches.

Regardless of the sequencing strategy and assembly method, we observed discrepancies between observed and expected abundances (**Figure 6**). Overall, the estimation based on MAGs obtained from Illumina data was not affected by the assembly approach, and similar levels of dispersion around the expected relative abundances were observed (±2%).

Regarding MAGs obtained with long-read sequencing technologies, we observed different trends depending on the assembly method. Assembly of PacBio data with metaFlye showed deviations from the expected abundances of approximately ±4%, with the only exception of *Staphylococcus epidermidis,* which exhibited a deviation of -10.5% (**Figure 6**). Using metaMDBG the estimated variance was around ±5%. We then considered whether the observed variance could be explained by the mean coverage depth of the mock community species. Both in PacBio data assembled with metaFlye and metaMDBG, we observed that the mean coverage depth was significantly correlated with the variance in species abundance estimates (-0.73 and -0.75, respectively). In particular, the negative correlation means that species with lower mean coverage depth tended to be more strongly underestimated (**Figure 6**).

Finally, species estimates derived from Nanopore data assembled with metaFlye were the closest to the expected values (**Figure 6**), except for *Clostridium beijerinckii* (- 30.0%) and *Escherichi coli* (-18.94%). Species abundance estimation based on MAGs obtained with metaMDBG assembled data showed a variation around ±5%, except for *Escherichia coli* (-8.4%). For both assembly approaches, we observed a relevant association with mean coverage depth, which was stronger for metaMDBG (-0.76) than for metaFlye (-0.56).

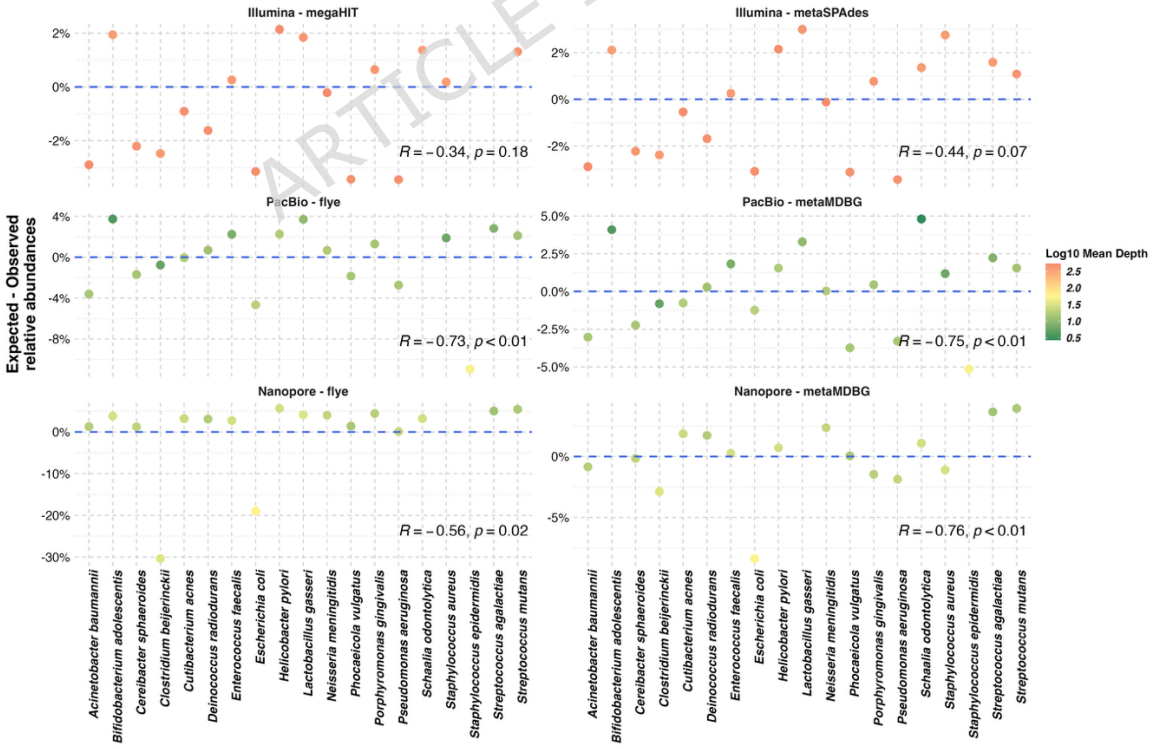

Figure 6: Comparison of relative abundances for mock community species*. Scatterplot representing the deviation between the estimated mock species abundances and the expected*

values. On the x-axis the mock species are reported and the estimated deviation on the y-axis (positive and negative values indicate deviations from the expected relative abundances). The dashed blue line indicates no deviation and the closer the dots are to the line, the more similar the observed and expected values are. Dot color shows the mean coverage depth (i.e., the number of times a specific base is covered) observed by mapping read to the reference genome. The Pearson r correlation between estimated variation and mean coverage depth and relative p-values, are shown in each panel at the bottom-right.

3.5 MAGs gene annotation

Medium- and high-quality MAGs were functionally annotated using Bakta. To avoid discrepancies due to different annotation pipelines, ATCC reference genomes were re-annotated using the same tool. Initially, the number of annotated gene types (i.e. CDS, tRNA, rRNA, ncRNA and tmRNA) were compared **(Figure 7)**. Overall, a comparable number of CDS were obtained in MAGs and ATCC reference genomes, with few differences. An underestimation of CDS was observed in 8 out of 20 species (namely *H. pylori*, *L. gasseri*, *N. meningitidis*, *P. gengivalis*, *S. aureus*, *S. epidermidis*, *S agalactiae*, and *S. mutans*) when MAGs obtained from short reads are considered, regardless of the assembly method **(Figure 7**). Regarding long reads, the number of predicted CDS was not influenced by assembly method in PacBio data, while for Nanopore data metaMDBG tended to produce less accurate annotation than metaFlye **(Figure 7**).

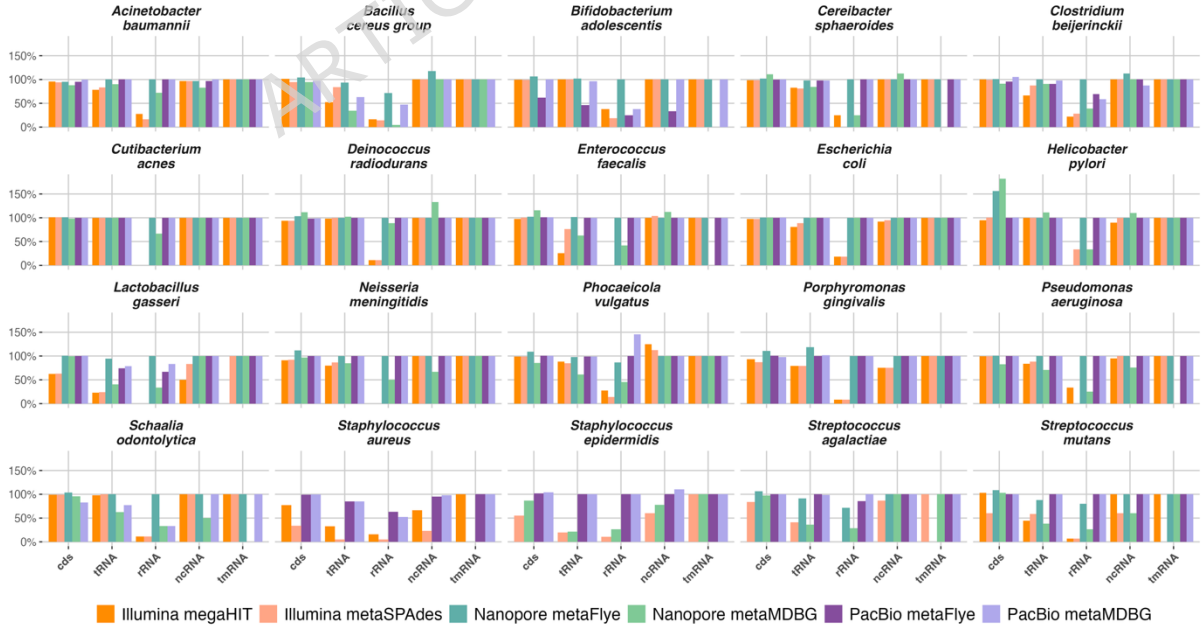

Figure 7: Amount of annotated genes for each type (CDS, tRNA, rRNA, ncRNA and tmRNA) for each genome. *On the y-axis the relative number of annotated genes in each MAGs compared with that expected in the reference genome is reported.*

Using MAGs retrieved from long reads, both Nanopore and PacBio, we observed an overall tendency to predict a number of ncRNA genes similar to those observed in reference genomes. Moreover, the number of predicted ncRNA genes in long-reads derived MAGs resulted influenced by assembly quality and reference genome coverage **(Figure 7)**. For instance, for MAGs obtained by assembling Nanopore data through metaMDBG and assigned to *C. beijerinckii* and *N. meningitidis* (both medium quality with a reference genome coverage ≤ 90%) an underestimation of both rRNA and tRNA genes was observed. Similarly, *S. odontolytica* MAGs obtained with PacBio data were the least accurate in terms of ncRNA genes annotation because of their lower completeness (73.12%, **Supplementary Table S5**). An underestimation of annotated ncRNA genes was observed in MAGs obtained from short reads, regardless of the assembly method. The impact of underestimation was associated with MAG completeness and gene redundancy (**Supplementary Table S5**). Indeed, for both Illumina derived MAGs classified as *L. gasseri* (both medium quality), the number of tRNA genes was underestimated compared to the reference genome and not rRNA genes were identified at all (**Supplementary Table S6**). Regarding gene redundancy, the number of genes for Alanine and Isoleucine tRNAs was underestimated in 14 out 20 species (**Supplementary Table S6**). Despite both *A. baumani* MAGs retrieved from short reads being classified as high quality, both were able to retrieve just 1 out of 7 expected Alanine tRNA genes (**Supplementary Table S6)**. A comparison of these seven genes demonstrated that 6 were identical while one was unique, sharing a 71% similarity with the others. The Alanine tRNA gene retrieved in both MAGs corresponded to the unique one in the reference genome (**Supplementary Table S6**).

We then investigated the presence of ribosomal RNA (rRNA) genes in the obtained MAGs. Using BLAST, we compared the 5S, 16S and 23S rRNA sequences annotated with Bakta in the ATCC reference genomes with those predicted in the MAGs to assess both the accuracy of rRNA gene assembly and prediction and the ability to recover intra-genomic variability. To reduce computational burden both the rRNA genes identified in ATCC genomes and MAGs were dereplicated (**Supplementary Table S7**). Unique genes were then compared using BLAST retaining only matches with at least 97% sequence identity and 90% query coverage (i.e. the unique rRNA gene annotated in the ATCC genome must be covered by the BLAST match for 90% of its sequence). For 16 out 20 species, no 16S and 23S genes were annotated

in MAGs obtained from Illumina assemblies (**Figure 7** and **Supplementary Table S7**). In *C. sphaeroides*, *C. acnes*, *H. pylori* and *P. gingivalis,* a single copy of both the 16S and 23S rRNA genes was annotated in Illumina derived MAGs, regardless of the assembly approach, whereas at least two copies were detected in the reference genomes (**Supplementary Table S7**), although all copies were identical. In MAGs obtained from PacBio data, at least one copy of both the 16S and 23S rRNA genes was detected. Furthermore, irrespective of the assembly approach, the number of 16S and 23S genes copies annotated in PacBio-derived MAGs was comparable to that observed in the reference genomes. When comparing unique 16S and 23S gene copies (i.e. gene sequences differing by at least 1 nt) we noticed differences among genomes (**Supplementary Table S7**). For instance, for *A. baumannii,* the intra-genomic 16S gene variability was recovered in MAGs, regardless of the assembly approach. On the contrary, regarding the 23S rRNA gene, the intra-genomic variability was caught only by metaMDBG. Finally, in Nanopore assembled MAGs, while the number of predicted 16S and 23S genes was comparable to the expected ones across the assembly approaches, the number of unique copies (i.e. intra-genomic variability) was not correctly identified. For instance, in *C. beijerinckii* 16S gene variability was overestimated in MAGs assembled with metaFlye (15) and underestimated with metaMDBG (5), compared to the expected value (13).

Regardless of the sequencing and assembly approaches, 5S genes were predicted in numbers comparable to those annotated in the reference genomes (**Supplementary Table S7**).

Furthermore, we also evaluated the quality of protein coding genes annotations by comparing the protein length profiles between MAGs and ATCC reference genomes (**Supplementary Figure S4**). Overall, no relevant differences were observed in the length profiles of MAGs retrieved from short reads compared to those of reference genomes, with the only exception of *L. gasseri* (medium quality MAGs). Considering MAGs obtained by binning metaMDBG and metaFlye contigs from Nanopore reads, we observed statistically relevant differences for 14 and 10 out of 20 species, respectively, 8 in common (namely C*. sphaeroides*, *D. radiodurans*, *H. pylori*, *N. meningitidis*, *P. vulgatus, S. odontolytica*, *S. agalactiae*, and *S.mutans*). When measuring the effect size as the mean difference in protein length among MAGs and reference genomes (d%), we observed values that generally ranged between -5% and 18% (**Supplementary Figure S4**), with worse performance in metaMDBG-assembled data and a tendency toward predicting

shorter CDS. For MAGs classified as *Helicobacter pylori* we observed the largest effect size, -31.3% and -52.7% in high and medium quality MAGs, assembled with metaFlye and metaMDBG, respectively. Regarding MAGs retrieved from PacBio we observed statically significant differences in protein length profiles in 4 out of 20 species (*B. cereus group*, *B. adolescentis*, *C. beijerinckii*, and *S. odontolytica*). Specifically, for all these 4 species metaMDBG MAGs were evaluated as medium quality with a negative effect size in protein length prediction, ranging between 5% and 10.7%. The only MAGs obtained with metaFlye (*C. beijerinckii*) reached a high-quality classification and the measured effect size (d%) was about -4.5%.

Finally, we evaluated the annotated protein genes by comparing the predicted functions in ATCC reference genomes to those in the obtained MAGs. A qualitative representation of the obtained results is available in **Figure 8**. PacBio-retrieved MAGs obtained the closest results compared to reference genomes (metaFlye Jaccard: mean 4.02%, median 0.65%, metaMDBG Jaccard: mean 4.39%, median 1.10%) with only three species showing relevant differences: *S. odontolytica* (metaMDBG, medium quality genome), *B. pacificus* (*B. cereus group*, metaMDBG) and *B. adolescentis* (metaFlye). These data were consistent with the observed reference genome coverage and assembly quality (**Supplementary Table S5**). Furthermore, three metaFlye assembled MAGs, classified as *S. mutans, P. aeruginosa* and *E. coli*, were the only ones to achieve a Jaccard distance from the reference equal to 0. Acceptable results were also obtained in Illumina retrieved MAGs (metaSPAdes Jaccard mean: 12.03% median: 4.22%; MEGAHIT Jaccard mean: 6.42% median: 3.70%). The species achieving the largest distance from the reference genomes were *Staphylococcus spp.*, *Streptococcus spp.* and *L. gasserii*. Nanopore data assembled with metaFlye produced results comparable to those obtained from short-reads (Jaccard mean: 8.73% median: 6.75%). Among these, *H. pylori* MAGs displayed the highest dissimilarity from the reference genomes (41.2%). In contrast, Nanopore MAGs assembled with metaMDBG were the most dissimilar to the reference genomes overall (Jaccard mean: 24.00% median: 26.05%).

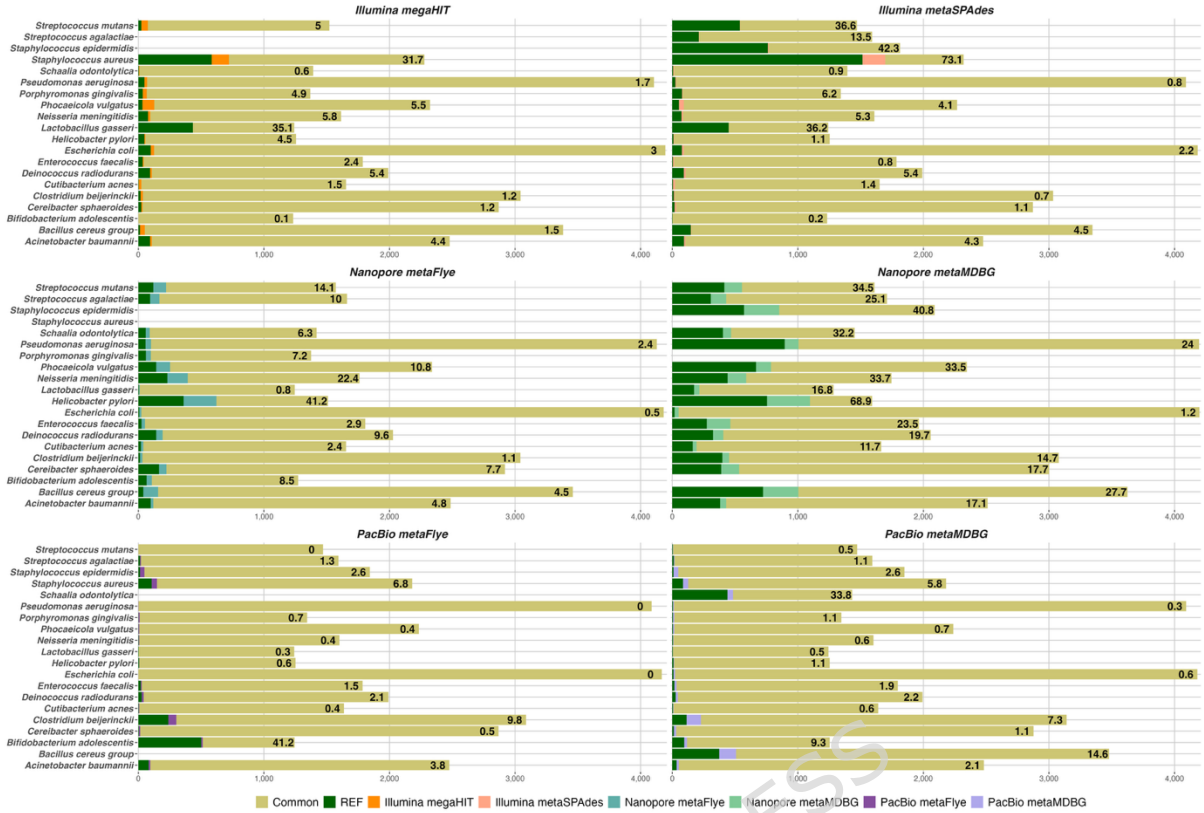

Figure 8: Stacked bar-plot showing the number of protein functions annotated in ATCC reference genome and in the corresponding MAGs. *Each panel corresponds to a combination of sequencing approach and assembly method. For each species the number of shared functions among the reference genome and the corresponding MAGs are shown in khaki, those exclusively found in reference genome in dark green and those peculiar to the exploited methodology in the color listed in the legend. In each bar the measured Jaccard distance among the reference genome and MAGs annotation is also reported as a percentage. The largest the value, larger is the discrepancy between protein annotations.*

4. DISCUSSION

Our understanding of the complex network between microorganisms and the surrounding environment, including the diversity, structure and dynamics of microbial communities, is still incomplete, due to the challenges that metagenomic studies face during library preparation, sequencing and analysis (9,57). Shotgun metagenomics is an informative approach to rapidly obtain compositional profiles of the investigated microbial community (i.e. reference based) or to retrieve nearly complete microorganism genomes through assembly-based approaches, including MAGs. The latter is gaining ever-growing interest in the research community, also due to decreasing sequencing costs (58). In this benchmark study, the impact of different sequencing technologies and metagenome assembler tools was thoroughly evaluated with the aim of assessing

high-quality MAG retrieval. Cutting-edge sequencing platforms, including the NovaSeq 6000 (Illumina), GridION (Nanopore) and Sequel IIe (PacBio) were employed in this work. The best-performing sequencer–assembler combination was assessed in terms of the overall assembly and MAG quality, species abundance/quantification estimation, concordance between reference and retrieved genomes, and functional annotation recovery. All the aspects discussed here provide an updated perspective compared with previous studies (23,31). A commercial mock community containing 20 bacterial species belonging to 18 genera was used, with a total DNA yield of 200 ng. The limited DNA yield and the presence of phylogenetically related species (i.e. congeneric taxa) reflect conditions that may occur naturally in biological samples. Moreover, this mock community is represented by a limited number of microorganisms allowing, a more thorough investigation of technical aspects.

For short-read sequencing (Illumina), a method involving chemical DNA fragmentation and PCR-based amplification steps during library preparation was used. As a result, Illumina generated the highest data output, reaching hundreds of millions of reads, complete genome coverage and the highest sequencing depth for all mock genomes. At the same time, we selected amplification-free protocols for long-read sequencing to align with the goal of single-molecule sequencing, producing reads longer than 10 kb, removing amplification bias while preserving base modifications (59). The ligation-sequencing protocol without a fragmentation step was chosen for Nanopore sequencing, whereas for the PacBio application, medium-sized fragments generated by mechanical fragmentation (with a peak size of approximately12 kbp) were sequenced. In both cases, the input DNA amount did not completely meet protocol requirements. Nonetheless, we obtained an adequate number of reads to perform bioinformatic analysis, with Nanopore producing a higher yield and longer reads compared to PacBio. PacBio HiFi sequencing, however, generated fewer reads overall but the trimming procedures did not affect their number, reflecting the inherently high base accuracy of HiFi data. This observation is consistent with recent literature showing that, although Nanopore can produce longer and more abundant reads, PacBio HiFi delivers highly accurate long reads that ultimately support superior genome reconstruction and higher-quality metagenome assemblies (39,60). The reduced number of reads obtained with PacBio sequencing probably affected the coverage and sequencing depth exclusively for the species *S. odontolytica,* compared to the other technologies. The strong impact of sequencing depth and uneven coverage on

MAG recovery observed for this low-abundance taxon is consistent with previous studies highlighting coverage as a major limiting factor for metagenome assembly and genome-resolved metagenomics (55, 59).

The effort in library preparation and to maximizing sequencing yield was crucial to avoid insufficient sequencing depth and coverage, which represent critical factors in metagenome reconstruction (39) and can influence the performance of meta-assembly tools (38). In fact, metagenome assembly is a crucial step for recovering near complete and high-quality MAGs. Two main algorithms have been used in this benchmark: de-Bruijn graph (61) and repeat graph (62). The first is principally used for short-read assembly, although application to third generation sequencing technologies are also available (39). The second is particularly suitable for long-read assembly. Here, two tools for each sequencing technology have been employed to consider the impact of the assembly step. For Illumina short-reads, the de-Bruijn graph-based tools MEGAHIT and metaSpades have been used, these differ in terms of algorithm heuristics and optimizations. Overall, metaSpades produced assemblies with larger N50 values and longer contigs than MEGAHIT. MetaSpades also enabled recovery of all the expected species with higher overall quality than MEGAHIT. In general, metaSpades performed slightly better and more precisely than MEGAHIT and also missed MAGs for two expected species (i.e. *S. agalactiae*, *S. epidermidis*). For long reads data, the repeat graph-based tool metaFlye and the de-Bruijn graph-based tool metaMDBG were employed. We chose two alternative approaches because it is well known that repeat graph-based approaches may be less effective at recovering low-abundance microorganisms and resolving strain-heterogeneity, which negatively impacts MAG quality (63). On the other hand, de Bruijn graph approaches rely on exact k-mers matching which is affected by the lower accuracy of long-reads compared to short-reads. MetaFlye on Nanopore and metaMDBG on PacBio assemblies reached the highest N50 values (~ 2Mbp) and coherently obtained the highest number of high-quality MAGs. This higher yield of high-quality MAGs from long-read data is consistent with previous HiFi- and Nanopore-based genome-resolved metagenomic studies, which showed that long and accurate reads substantially increase the number and quality of recovered MAGs compared to short-read assemblies (37). Furthermore, although Nanopore with both assemblers and PacBio with metaFlye missed MAGs for some expected species, metaMDBG on PacBio retrieved MAGs for all the mock community species. Thus, the Illumina-metaSpades and PacBio-metaMDBG combinations allow better resolution of

biodiversity by distinguishing co-generic species. On the contrary, both Illumina-MEGAHIT and Nanopore-metaMDBG fail to discriminate between congeneric genomes. In terms of assembly contiguity, the combination of PacBio and metaMDBG obtained the highest number of reference genomes (i.e. 12) covered by the lowest number of contigs, followed by Nanopore and metaFlye (i.e. 8) (**Figure 3**). This demonstrates that the repeat-graph based assembler metaFlye can compensate for the lower read quality and uneven coverage typical of Nanopore data, while the sparse de Bruijn graph-based assembler metaMDBG performs better with long high-quality reads generated by PacBio. In line with earlier platform comparisons, third-generation sequencing generally improves assembly contiguity compared to Illumina short reads (55).

When comparing the obtained MAGs with their reference genomes using sketch-based distances, higher MAG quality corresponds to lower measured distances. Dereplication of obtained MAGs and reference ATCC genomes through ANI estimation produced intriguing results with 13 out of 20 best genomes chosen among MAGs generated from long read data (**Figure 4, Supplementary Table S4**). This is a further demonstration that the recovered MAGs, particularly those obtained with the PacBio-metaMDBG, can be at least as representative as their reference genome. Nonetheless, MAG sizes were compared with those of the reference genomes (**Supplementary Figure S2**). Overall, Illumina data tended to not accurately estimate genome sizes, whereas assembler choice influenced Nanopore assemblies. PacBio data were more consistent with the reference genomes and less affected by the assembly approach. Regardless of sequencing technology and assembly approach, *D. radiodurans* and *R. sphaeroides* genome sizes were overestimated. This finding could be due to the overall quality of the reference genomes, which are both fragmented. In fact, in **Figure 4** for these two species, two PacBio-metaFlye MAGs were chosen as centroid dereplication by dRep.

Furthermore, we also evaluated the ability of MAGs obtained with different sequencing technologies to recapitulate the species abundances in the mock community. Nanopore data assembled with metaFlye provided the closest estimates to the expected species abundances, with the only exception of *E. coli*, the genome with the highest coverage depth in the dataset. Similar trends were observed for PacBio data, demonstrating that, while genome efficient

reconstruction can be achieved with different assembly approaches, the estimation of species abundances is affected by sequencing depth.

The key novelty of this work lies in the comprehensive evaluation of how different sequencing technologies and combined assembly strategies influence genome annotation, specifically in terms of the number of annotated genes, length of predicted proteins, and number and types of inferred functions. First, regarding gene prediction, short reads and Nanopore coupled with metaMDBG MAGs tended to underestimate CDS and non-coding genes (**Figure 7**), principally due to MAG incompleteness and gene redundancy. Nonetheless, PacBio-derived MAGs, as well as Nanopore-metaFlye MAGs, allowed the annotation of all copies of redundant genes such as alanine tRNA genes. It is worth emphasizing the importance of accurately assembling and annotating both ncRNA (e.g. sRNA) and tRNA in prokaryotes, considering their essential role in gene expression regulation and responding to environmental stressors (64). Furthermore, ncRNAs have been described as key regulator of virulence genes in human pathogens, such as *Staphylococcus aureus* (65). Moreover, our data confirms the difficulties in reconstructing rRNA genes by using short reads (66), because these genes are highly repetitive and because De Brujin graph-based approaches, may collapse nearly identical or in tandem repeated operons (67). By contrast, in long-read derived MAGs it is possible to retrieve multiple rRNAs copies and capture rRNA intra-genomic variability, as already demonstrated with targeted approaches (5,68). Second, regarding protein length distributions, **Supplementary Figure S4** shows a statistically significant underestimation of protein length for MAGs from long-reads, particularly for Nanopore-metaMDBG (14 MAGs) and metaFlye, (10 MAGs) and PacBio-metaMDBG (4 MAGs) combinations. This highlighted the impact of MAGs completeness in functional annotation. Furthermore, also the effect size (d%) of the evaluated differences could be explained by MAGs quality and assembly contiguity. Indeed, MAGs obtained by assembling Nanopore data were those with the lowest N50 (**Supplementary Table S5**) and this may explain the observed decrease in mean protein length (**Supplementary Figure S4**). It is worth highlighting that Liu et al., applied five rounds of genome polishing to retrieve highly accurate MAGs, after assembling ONT reads with metaFlye and binning with metaWRAP (69). This is a bioinformatic procedure aimed at reducing missassemblies by comparing contigs with raw reads. Indeed, as we demonstrated in our results, and as already reported in literature, ONT assemblies are less accurate compared to PacBio assemblies (62). To reduce the impact of

missassemblies in genomic projects, several rounds of polishing are suggested, possibly also using short reads (70). As demonstrated by Liu et al., applying polishing on metagenomic data becomes computationally expensive and for this reason we decided not to include this step in our analysis, which may explainthe lower accuracy in gene annotation in Nanopore data compared to other technologies. Finally, when comparing inferred functions between reference genomes and MAGs, Nanopore-metaMDBG combination was the worst-performing, followed by Illumina, whereas Nanopore-metaFlye and PacBio-metaMDBG showed similar performances. By contrast, PacBio-metaFlye showed the best performance and perfect concordance for 3 MAGs.

5. CONCLUSIONS

High-throughput sequencing technologies have substantially advanced microbial ecology, with ongoing developments in shotgun metagenomics, assembly methods, and long-read sequencing improving genome reconstruction (3,4,14,55, 69). A major goal in the field is to obtain a comprehensive understanding of microbial diversity, genomic function and microbial interactions. In this study, we compared current metagenomic sequencing strategies to evaluate their performance and limitations, highlighting how factors such as sequencing platform, depth, read length, and assembly algorithms can influence genome reconstruction.

Although additional validation across a wider range of samples and more complex environments will be required to strengthen and generalize our observations, our findings provide useful insights for future methodological advancements aimed at improving strain-resolved metagenome assembly.

Overall, our results indicate that these methodological choices can lead to measurable differences in genome reconstruction outcomes, underscoring the importance of selecting appropriate strategies according to the specific goals of the study. In this context, our benchmark provides a practical framework for selecting the most appropriate combination of sequencing platform and assembly strategy.

ABBREVIATIONS

HTS	high throughput sequencing technologies
MAGs	Metagenome-assembled Genomes
ANI	Average Nucleotide Identity
CDS	protein coding DNA sequencing
tRNA	transfer RNA
rRNA	ribosomal RNA
tmRNA	transfer-messenger RNA
ncRNA	non-coding RNA

SUPPLEMENTARY INFORMATION

Supplementary Figure S1: Genomic profile of the commercial mock microbial community (MSA-1002, ATCC) obtained from Femto Pulse System with Genomic DNA 165 kb Kit. (**A**) The Femto Pulse measurement of No-fragmented mock genomes showed a smear from 1 kbp to 2Mbp. (**B**) Femto Pulse profile of mock genomes after shearing by using Megaruptor3 displayed 99% of the fragments between 1-30 kb, with a peak of about 12 kbp.

Supplementary Figure S2: **Comparison of the MAGs size to the expected genomes length for each combination of sequencing technology and assembly approach.** Each dot represents the longest contig and is colored according to coverage value.

Supplementary Figure S3: Relative Abundances of Mock community species. The relative species abundances estimated for the 20-genome mock community (*Ref*) and for the MAGs are shown as stacked barplots.

Supplementary Figure S4: Boxplot showing the proteins length distribution profile. A panel per Mock species is drawn. Pairwise comparisons were performed by using the Wilcoxon test and protein length profile from the corresponding ATCC genome as reference. The obtained p-values were adjusted by Holm-Bonferroni method. Exclusively for relevant comparisons the level of significance using stars (* p-value ≤ 0.05, ** p- value ≤ 0.01, *** p- value ≤ 0.001, **** p- value ≤ 0.0001) and the change between MAGs and reference proteins mean lengths percent (d%) are shown.

Supplementary Table S1: Taxonomic composition and genomic characteristics of the Mock ATCC 20 Strain Even Mix (MSA-1002, ATCC, USA). Per each strain, the species scientific name, the genome size and the estimated relative abundance are listed. The last version of the Product Sheet was available on the ATCC website in 2022.

Supplementary Table S2: Mapping results for Illumina, Nanopore and PacBio reads against mock reference genomes. Each row represents a contig belonging to its reference genome. This table reports number of mapped reads (*n. reads*), covered bases (*cov. bases*), coverage percentages (*coverage*), mean depth *(meandepth*), mean base quality *(meanbaseq*), mean map quality (*meanmapq*), Read Per Kilobase per Milion (RPKM).

Supplementary Table S3: Mapping results for Illumina, Nanopore and PacBio contigs obtained by related meta-assemblers against mock reference genomes. Each row represents a reference genome. For each combination of sequencing technology and meta-assembler, this table reports the number of mapped contigs (*Fragments*) and the covered bases percentage (*Coverage*).

Supplementary Table S4: Results of kMetaShot classification of each MAG. The dereplication results of MAGs and reference genomes obtained using dRep are also reported in the columns named *primary_cluster, secondary_cluster and Representative Genome.*

Supplementary Table S5: Quality check of MAGs. Completeness and contamination percentages for each bin/MAG, assessed with CheckM, are reported. Also, MAG-specific N50 and GC% are also computed. Furthermore, total length of each MAG is compared with the length of the expected genome relying on kMetaShot classification and expressed in percentage in the last column (*% covered*).

Supplementary Table S6: Number of aminoacidic specific tRNA genes annotated for MAG and reference genomes (in red).

Supplementary Table S7: Comparison of the number annotate ribosomal RNA genes in each of the 20 mock species both in ATCC reference genomes and assigned MAGS (in red) and their comparison through blast. Per each species is shown the comparison between the number of predicted ribosomal rRNA genes (i.e. 5S,16S, and 23S) in reference genomes and MAGs. In each line the species name (Species), combination of sequencing and assembly approach (Approach), the rRNA gene (rRNA), the number of predicted genes in reference genomes (Expected total), the number of unique copies in reference genomes (Expected unique), the number of predicted genes in MAG (MAG total), the number of unique copies in MAG (MAG unique) and the number of match(es) obtained by comparing unique genes between those predicted in ATCC genomes and those in MAGs (Match) by using blast (Blast results were filtered according to identity percentage ≥ 97% and query coverage ≥ 90%).

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated during the current study are available in the ENA repository, reference number BioProject PRJEB89875 (https://www.ebi.ac.uk/ena/browser/view/PRJEB89875). Command-line used for data analysis are available as GitHub repository at https://github.com/bfosso/MAGs_paper.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by projects: Life Science Hub Regione Puglia (LSH-Puglia, T4-AN-01 H93C22000560003), INNOVA - Italian network of excellence for advanced diagnosis (PNC-EJ-2022-23683266 PNC-HLS-DA), DARE - DigitAl lifelong pRevEntion initiative (PNC-I.1 "Research initiatives for innovative technologies and

pathways in the health and welfare sector" D.D. 931 of 06/06/2022, code PNC0000002, CUP: B53C22006420001), and by ELIXIR-IT through the PNRR Project ELIXIRxNextGenIT - ELIXIR x NextGenerationIT: consolidation of the Italian Infrastructure for Omics Data and Bioinformatics (Grant Code IR0000010, CUP:B53C22000690005).

Authors' contributions

Conceptualization, G.V., E.N., B.F., M.M. and G.P.; methodology, G.V., E.N., G. D, M.F.C., B.F. and M.M.; validation, G.V., E.N., G. D, B.F. and M.M.; formal analysis, G.V., E.N., G. D, B.F. and M.M.; investigation, G.V., E.N., G. D, B.F. and M.M.; resources, G.P.; data curation, G.V., E.N., G. D, B.F. and M.M.; writing—original draft preparation, G.V., E.N., G. D, B.F. and M.M. writing—review and editing, G.V., E.N., G. D, M.F.C., S.N.C, B.F., M.M. and G.P.; visualization, E.N., G.V. and G.D.; supervision, G.P.; project administration, B.F. and M.M.; funding acquisition, G.P. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

Not applicable

REFERENCES

1. Staley JT, Konopka A. MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS. Annu Rev Microbiol. 1985 Oct;39(1):321–46.
2. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microbial Genomics [Internet]. 2020 Aug 1 [cited 2022 Nov 24];6(8). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000409
3. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics. 2021 Jan 18;22(1):178–93.
4. Purushothaman S, Meola M, Egli A. Combination of Whole Genome Sequencing and Metagenomics for Microbiological Diagnostics. IJMS. 2022 Aug 30;23(17):9834.
5. Notario E, Visci G, Fosso B, Gissi C, Tanaskovic N, Rescigno M, et al. Amplicon-Based Microbiome Profiling: From Second- to Third-Generation Sequencing for Higher Taxonomic Resolution. Genes. 2023 Jul 31;14(8):1567.

6. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011 Dec;480(7376):241–4.
7. Li C, Chen J, Li SC. Understanding Horizontal Gene Transfer network in human gut microbiota. Gut Pathog. 2020 Dec;12(1):33.
8. Jiang Y, Wang Y, Che L, Yang S, Zhang X, Lin Y, et al. GutMetaNet: an integrated database for exploring horizontal gene transfer and functional redundancy in the human gut microbiome. Nucleic Acids Research. 2025 Jan 6;53(D1):D772–82.
9. Carabeo-Pérez A, Guerra-Rivera G, Ramos-Leal M, Jiménez-Hernández J. Metagenomic approaches: effective tools for monitoring the structure and functionality of microbiomes in anaerobic digestion systems. Appl Microbiol Biotechnol. 2019 Dec;103(23–24):9379–90.
10. Silverstein MR, Segrè D, Bhatnagar JM. Environmental microbiome engineering for the mitigation of climate change. Global Change Biology. 2023 Apr;29(8):2050–66.
11. Anyansi C, Straub TJ, Manson AL, Earl AM, Abeel T. Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. Front Microbiol. 2020 Aug 18;11:1925.
12. Lapidus AL, Korobeynikov AI. Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. Front Microbiol. 2021 Mar 23;12:613791.
13. Pinto Y, Bhatt AS. Sequencing-based analysis of microbiomes. Nat Rev Genet. 2024 Dec;25(12):829–45.
14. Kim N, Ma J, Kim W, Kim J, Belenky P, Lee I. Genome-resolved metagenomics: a game changer for microbiome medicine. Exp Mol Med. 2024 Jul 1;56(7):1501–12.
15. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. bioRxiv [Internet]. 2019 Settembre [cited 2019 Dec 12]; Available from: http://biorxiv.org/lookup/doi/10.1101/762302
16. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023 Feb 23;1–12.
17. The Genome Standards Consortium, Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017 Aug;35(8):725–31.
18. Defazio G, Tangaro MA, Pesole G, Fosso B. kMetaShot: a fast and reliable taxonomy classifier for metagenome-assembled genomes. Briefings in Bioinformatics. 2024 Nov 22;26(1):bbae680.
19. Sabina J, Leamon JH. Bias in Whole Genome Amplification: Causes and Considerations. In: Kroneis T, editor. Whole Genome Amplification [Internet]. New York, NY: Springer New York; 2015 [cited 2025 Jan 9]. p. 15–41. (Methods in Molecular Biology; vol. 1347). Available from: https://link.springer.com/10.1007/978-1-4939-2990-0_2
20. Nelson MT, Pope CE, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, et al. Human and Extracellular DNA Depletion for Metagenomic Analysis of

Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. Cell Reports. 2019 Feb;26(8):2227-2240.e5.
21. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn LJ, et al. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. Front Microbiol. 2019 Jun 12;10:1277.
22. McArdle AJ, Kaforou M. Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. Access Microbiology [Internet]. 2020 Apr 1 [cited 2025 Jan 9];2(4). Available from: https://www.microbiologyresearch.org/content/journal/acmi/10.1099/acmi.0.000104
23. Latorre-Pérez A, Pascual J, Porcar M, Vilanova C. A lab in the field: applications of real-time, in situ metagenomic sequencing. Biology Methods and Protocols. 2020 Jan 1;5(1):bpaa016.
24. Kim HJ, Ahn DH, Yu Y, Han H, Kim SY, Joo JY, et al. Microbial profiling of peri-implantitis compared to the periodontal microbiota in health and disease using 16S rRNA sequencing. J Periodontal Implant Sci. 2023;53(1):69.
25. Arredondo A, Àlvarez G, Isabal S, Teughels W, Laleman I, Contreras MJ, et al. Comparative 16S rRNA gene sequencing study of subgingival microbiota of healthy subjects and patients with periodontitis from four different countries. J Clinic Periodontology. 2023 Sep;50(9):1176–87.
26. Marzano M, Fosso B, Colliva C, Notario E, Passeri D, Intranuovo M, et al. Farnesoid X receptor activation by the novel agonist TC-100 (3α, 7α, 11β-Trihydroxy-6α-ethyl-5β-cholan-24-oic Acid) preserves the intestinal barrier integrity and promotes intestinal microbial reshaping in a mouse model of obstructed bile acid flow. Biomedicine & Pharmacotherapy. 2022 Sep;153:113380.
27. Tumolo M, Salerno C, Manzari C, Vergine P, Marzano M, Notario E, et al. Linking feed, biodiversity, and filtration performance in a Self-Forming Dynamic Membrane BioReactor (SFD MBR) treating canning wastewater. Journal of Water Process Engineering. 2024 Sep;66:106031.
28. Roy G, Prifti E, Belda E, Zucker JD. Deep learning methods in metagenomics: a review. Microbial Genomics [Internet]. 2024 Apr 17 [cited 2025 Mar 19];10(4). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001231
29. Ben Khedher M, Ghedira K, Rolain JM, Ruimy R, Croce O. Application and Challenge of 3rd Generation Sequencing for Clinical Bacterial Studies. IJMS. 2022 Jan 26;23(3):1395.
30. Govender KN, Eyre DW. Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. Microbial Genomics [Internet]. 2022 Oct 31 [cited 2025 May 15];8(10). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000886

31. Meslier V, Quinquis B, Da Silva K, Plaza Oñate F, Pons N, Roume H, et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data. 2022 Nov 11;9(1):694.
32. Bokulich NA, Ziemski M, Robeson MS, Kaehler BD. Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. Computational and Structural Biotechnology Journal. 2020;18:4048–62.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
34. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics. 2015 May 15;31(10):1674–6.
35. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017 May;27(5):824–34.
36. Leger A, Leonardi T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. JOSS. 2019 Feb 28;4(34):1236.
37. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Zhang Z, editor. Bioinformatics Advances. 2023 Jan 5;3(1):vbac085.
38. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020 Nov;17(11):1103–10.
39. Benoit G, Raguideau S, James R, Phillippy AM, Chikhi R, Quince C. High-quality metagenome assembly from long accurate reads with metaMDBG. Nat Biotechnol. 2024 Sep;42(9):1378–83.
40. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011 May 2;17(1):10.
41. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
42. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. Zou Q, editor. PLoS ONE. 2016 Oct 5;11(10):e0163962.
43. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018 Dec;6(1):158.
44. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019 Jul 26;7:e7359.
45. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016 Feb 15;32(4):605–7.
46. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. CONCOCT: Clustering cONtigs on COverage and ComposiTion [Internet]. arXiv; 2013 [cited 2025 May 19]. Available from: https://arxiv.org/abs/1312.4038

47. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043–55.
48. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 Dec;17(1):132.
49. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Borgwardt K, editor. Bioinformatics. 2022 Nov 30;38(23):5315–6.
50. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. The ISME Journal. 2017 Dec 1;11(12):2864–8.
51. Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 1963 Mar;58(301):236–44.
52. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microbial Genomics [Internet]. 2021 Nov 30 [cited 2025 Apr 10];7(11). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685
53. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014 Nov;11(11):1144–6.
54. Mirete S, Sánchez-Costa M, Díaz-Rullo J, González de Figueras C, Martínez-Rodríguez P, González-Pastor JE. Metagenome-Assembled Genomes (MAGs): Advances, Challenges, and Ecological Insights. Microorganisms. 2025 Apr 25;13(5):985.
55. Hördt A, López MG, Meier-Kolthoff JP, Schleuning M, Weinhold LM, Tindall BJ, et al. Analysis of 1,000+ Type-Strain Genomes Substantially Improves Taxonomic Classification of Alphaproteobacteria. Front Microbiol. 2020 Apr 7;11:468.
56. Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T, Kyrpides NC, et al. Genome-Based Taxonomic Classification of the Phylum Actinobacteria. Front Microbiol. 2018 Aug 22;9:2007.
57. Filardo S, Di Pietro M, Sessa R. Current progresses and challenges for microbiome research in human health: a perspective. Front Cell Infect Microbiol. 2024 Apr 4;14:1377012.
58. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020 Sep;18(9):491–506.
59. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Dec;21(1):30.

60. Bein B, Chrysostomakis I, Arantes LS, Brown T, Gerheim C, Schell T, et al. Long-read sequencing and genome assembly of natural history collection samples and challenging specimens. Genome Biol. 2025 Feb 10;26(1):25.
61. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov;29(11):987–91.
62. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.
63. Bickhart DM, Kolmogorov M, Tseng E, Portik DM, Korobeynikov A, Tolstoganov I, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol. 2022 May;40(5):711–9.
64. Majumder R, Ghosh S, Das A, Singh MK, Samanta S, Saha A, et al. Prokaryotic ncRNAs: Master regulators of gene expression. Current Research in Pharmacology and Drug Discovery. 2022 Jan 1;3:100136.
65. Desgranges E, Marzi S, Moreau K, Romby P, Caldelari I. Noncoding RNA. Microbiol Spectr. 7(2):10.1128/microbiolspec.gpp3-0038–2018.
66. Zhang Z, Xiao J, Wang H, Yang C, Huang Y, Yue Z, et al. Exploring high-quality microbial genomes by assembling short-reads with long-range connectivity. Nat Commun. 2024 May 31;15(1):4631.
67. Mise K, Iwasaki W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. ISME Commun. 2022 Dec 1;2(1):118.
68. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov 6;10(1):5029.
69. Liu L, Yang Y, Deng Y, Zhang T. Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. Microbiome. 2022 Dec 2;10(1):209.
70. Luan T, Commichaux S, Hoffmann M, Jayeola V, Jang JH, Pop M, et al. Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. BMC Genomics. 2024 Jul 8;25(1):679.
71. Hoang MTV, Irinyi L, Hu Y, Schwessinger B, Meyer W. Long-Reads-Based Metagenomics in Clinical Diagnosis With a Special Focus on Fungal Infections. Front Microbiol. 2022 Jan 6;12:708550.