nature methods

Review article

https://doi.org/10.1038/s41592-024-02262-1

Unveiling microbial diversity: harnessing long-read sequencing technology

Received: 8 September 2022

Accepted: 29 March 2024

Published online: 30 April 2024

Check for updates

Daniel P. Agustinho[1], Yilei Fu[2], Vipin K. Menon[1,3], Ginger A. Metcalf[1], Todd J. Treangen[2,4] & Fritz J. Sedlazeck[1,2]

Long-read sequencing has recently transformed metagenomics, enhancing strain-level pathogen characterization, enabling accurate and complete metagenome-assembled genomes, and improving microbiome taxonomic classification and profiling. These advancements are not only due to improvements in sequencing accuracy, but also happening across rapidly changing analysis methods. In this Review, we explore long-read sequencing's profound impact on metagenomics, focusing on computational pipelines for genome assembly, taxonomic characterization and variant detection, to summarize recent advancements in the field and provide an overview of available analytical methods to fully leverage long reads. We provide insights into the advantages and disadvantages of long reads over short reads and their evolution from the early days of long-read sequencing to their recent impact on metagenomics and clinical diagnostics. We further point out remaining challenges for the field such as the integration of methylation signals in sub-strain analysis and the lack of benchmarks.

High-throughput sequencing technologies were first used in metagenomic studies in 2006 (ref. 1), initiating a transformative paradigm shift within the field. Over time, these technologies have become more cost-efficient and widely utilized, profoundly impacting metagenomics. Another crucial advancement emerged in the early 2010s with the introduction of long-read sequencing, ushering in a new era of metagenomic exploration. Compared to short-read technologies, long-read technologies (Table 1) allow for the sequencing of both intergenomic and intragenomic repetitive regions and their neighboring sequences in the same read, resulting in less fragmented metagenome assembled genomes (including plasmids), more accurate taxonomic characterization (species/strain level)[2], improved detection of horizontal gene transfer[3], and cataloging of large structural variations[4,5], such as duplications or inversions. Long reads can also be used for identifying methylation patterns, as they can detect modified nucleotides during sequencing, and the reads are long enough to map full methylation sites, especially in repetitive regions[6]. Nevertheless, long reads also have drawbacks, such as increased costs and DNA amount requirements[7]. Furthermore, in the context of metagenomics, DNA extraction often yields shorter fragments, which might negate the full benefits of long-read sequencing.

Recently, long-read sequencing has matured considerably, providing continuous cost and yield improvements, and reduced sequencing error rates (Table 1). In addition, there have also been reduced input DNA requirements, allowing the sequencing of smaller sample DNA concentrations. Currently, two main companies dominate the long-read market: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio delivers very highly accurate long reads, while ONT produces slightly less accurate reads, but with multiple devices operating at different scales, from handheld sequencing devices (MinION) to high-throughput sequencers (PromethION). To date, ONT is the only technology capable of performing direct RNA sequencing (RNA-seq)[8], although it is still in the experimental stage (Fig. 1a).

In this Review, we examine the application of long-read sequencing in metagenomics, encompassing both targeted and non-targeted strategies. We evaluate contemporary analytical methodologies that facilitate more comprehensive metagenomic analyses, assessing diverse

[1]Human Genome Sequencing center, Baylor College of Medicine, Houston, TX, USA. [2]Department of Computer Science, Rice University, Houston, TX, USA. [3]Senior research project manager, Human Genetics, Genentech, South San Francisco, CA, USA. [4]Department of Bioengineering, Rice University, Houston, TX, USA. e-mail: fritz.sedlazeck@bcm.edu

Table 1 | Comparison between short-read and long-read technologies

Platform	MiSeq	NovaSeq 6000	Sequel II	Revio	Flongle	MinION	GridION	PromethION
Company	Illumina	Illumina	PacBio	PacBio	Nanopore	Nanopore	Nanopore	Nanopore
Read length (average)[a]	Up to 2×300 bp	Up to 2×250 bp[b]	~13.5–20 kb[52]	15–18 kb	20 kb	20 kb	20 kb	20 kb
Yield per cell (Gb)	Up to 15	~350	25 (HiFi)	90	1–1.5	15–20	15–20	~120
Runtime (h)	4–56	13–44	30	24	16–24	72	72	72
Read accuracy (Q20+)	99.75%[128]	99.75%[128]	99.18–99.8%[52]	>99.5%[129]	97–99%[12]	97–99%[12]	97–99%[12]	97–99%[12]
Can perform Direct RNA-seq	No	No	No	No	Yes	Yes	Yes	Yes
DNA input needed	1–500 ng[18]	1–500 ng[18]	150 ng–1 µg[7] [c]	150 ng–1 µg[7] [c]	150 ng–1 µg[7]	150 ng–1 µg[7]	150 ng–1 µg[7]	150 ng–1 µg[7]
Estimated sequencing costs per Gb[d]	US$178–1,705[130,131]	US$3.95[131]	US$30–43[131,132]	US$8–11[132]	US$118–437[131,132]	US$21–51[131,132]	US$29–51[132]	US$6–12[132]

[a] Complex metagenomic samples very rarely reach average lengths superior to 2.5–9.0 kb (see text for details). [b] Available only on Illumina SP flow cells. [c] An exception might be the ultra-low-input protocol for PacBio. [d] Please note that these sequencing costs are estimates, subject to rapid changes, and the information presented here may quickly become outdated.

approaches, differentiating between mapping and assembly-based methods based on their respective utilities with long reads. Additionally, we not only scrutinize recent progress but also illuminate existing constraints and potential future prospects within the field. This Review is intended to provide researchers and practitioners with a comprehensive overview of the latest trends and computational techniques in the field, empowering them to leverage long-read technologies and methodologies to drive discoveries and push the boundaries of metagenomics research.

Long-read sequencing technologies in the context of metagenomics

Illumina short-read technology is still widely used and has great advantages such as low cost, high throughput, accuracy and sensitivity across low-abundance genomes, but also has its limitations[5,7]. Repetitive genomic regions or highly homologous regions can impact the efficiency of assembly and alignment tools, especially if these regions are longer than the overlapping length of reads or contigs[9]. Consequently, complete genome resolution using short reads is nearly always unattainable[10]. In metagenomic approaches, taxonomic differentiation in the sample can become complicated as there are probably multiple closely related strains whose genomes differ in just a few locations[11]. These intergenomic repeats, segments shared by various organisms, are difficult to resolve computationally without simultaneously sequencing their neighboring regions[4,5,7]. One potential strategy to overcome these issues is to increase the read length, thus increasing the likelihood that a read includes an organism's specific region to improve identification[9].

Longer reads produced by ONT and PacBio aim to fill this gap. PacBio's single-molecule, real-time (SMRT) technology yields high-fidelity (HiFi) long reads at a very low error rate[9]. Each SMRT Cell (8 M) for the PacBio Sequel II system can generate up to 80 Gb of sequence data in ~30 h, with an average read length of ~15–20 kb, or about 4 million reads. The new PacBio Revio system can run up to four SMRT Cells (25 M) in parallel, with each producing up to 90 Gb of data for a total of 360 Gb in ~24 h (Table 1). PacBio library preparation generally takes a minimum of 7 h.

ONT library preparation generally takes 1–2 h, and the sequencing runs last ~72 h for most of the platforms used, except for Flongle (16–24 h; Table 1). The MinION and PromethION platforms have an average read length of 13–20 kb and can reach up to 4 Mb[12], which would be sufficient to encompass whole chromosomes/genomes for some organisms. ONT machines' yield varies according to the kits and platforms used[12], and it can range from up to 2.8 Gb (Flongle) passing through the ~10–20 Gb range (MinION), all the way up to ~100 Gb (PromethION R10, kit14). This newest ONT technology allows for reads that are slightly less accurate than short reads (~1–3% sequencing error)[12]. ONT long reads are in general a bit less accurate in their base calling than PacBio HiFi reads but have a higher throughput and lower cost[7]. One of the biggest advantages of the ONT MinION and Flongle platforms is their portability due to the small size and off-line functionality of the machines, allowing them to be taken on expeditions to isolated places and performing DNA sequencing in situ and in isolated regions such as the International Space Station[13], for example.

In addition to nucleotides sequences, long reads can simultaneously read out the methylation status on CpG islands for 5mC and 5hmC[6] (more details below). Most interestingly, ONT has the ability to read this out not only on DNA, but also directly via sequencing RNA[8]. This allows further separation of different organisms within a sample (Fig. 1a).

Thus, while long-read sequencing is still not as common as short-read sequencing, it holds a remarkable promise with ever-decreasing costs and error rates. Nevertheless, it is worth noting that both platforms (ONT and Pacbio) rely on the quality and quantity of the input DNA. Overall, the use of long-read-based metagenomics is developing rapidly (Fig. 1b). Novel analytical approaches and bioinformatic tools are constantly being developed to fully exploit their properties. Their importance has been shown in the clinical context as well, such as the surveillance of antibiotic resistance genes[14] and pathogens in the hospital environment and potential hospital-associated infections[15], and in the identification of recent outbreaks, such as monkeypox[16] and coronavirus disease 2019 (COVID-19)[17].

The success of a metagenomic study hinges on various factors, starting with the initial experimental design, including the selection of technology and whether a targeted or non-targeted approach is used. Notably, long-read sequencing typically needs a larger quantity of high-quality DNA compared to short-read studies. PacBio- and ONT-specific extraction can require a minimum input of 150 ng up to 1 µg of amplification-free, high-quality genetic material for sequencing single-organism samples[7]. In comparison, Illumina protocols normally need as low as 1–500 ng of DNA[18]. Table 1 shows a comparison across the technologies.

Extracting genetic material from complex communities can be a challenging process due to the variety of organisms found, which may each require specific extraction protocols (for example, circularized or linear DNA). As a result, distinct DNA fragments are generated for different organisms, limiting the average read length obtained in such studies to around 2.5–9.0 kb for both synthetic and real microbial communities[19]. Organisms with low biomass might not be represented in

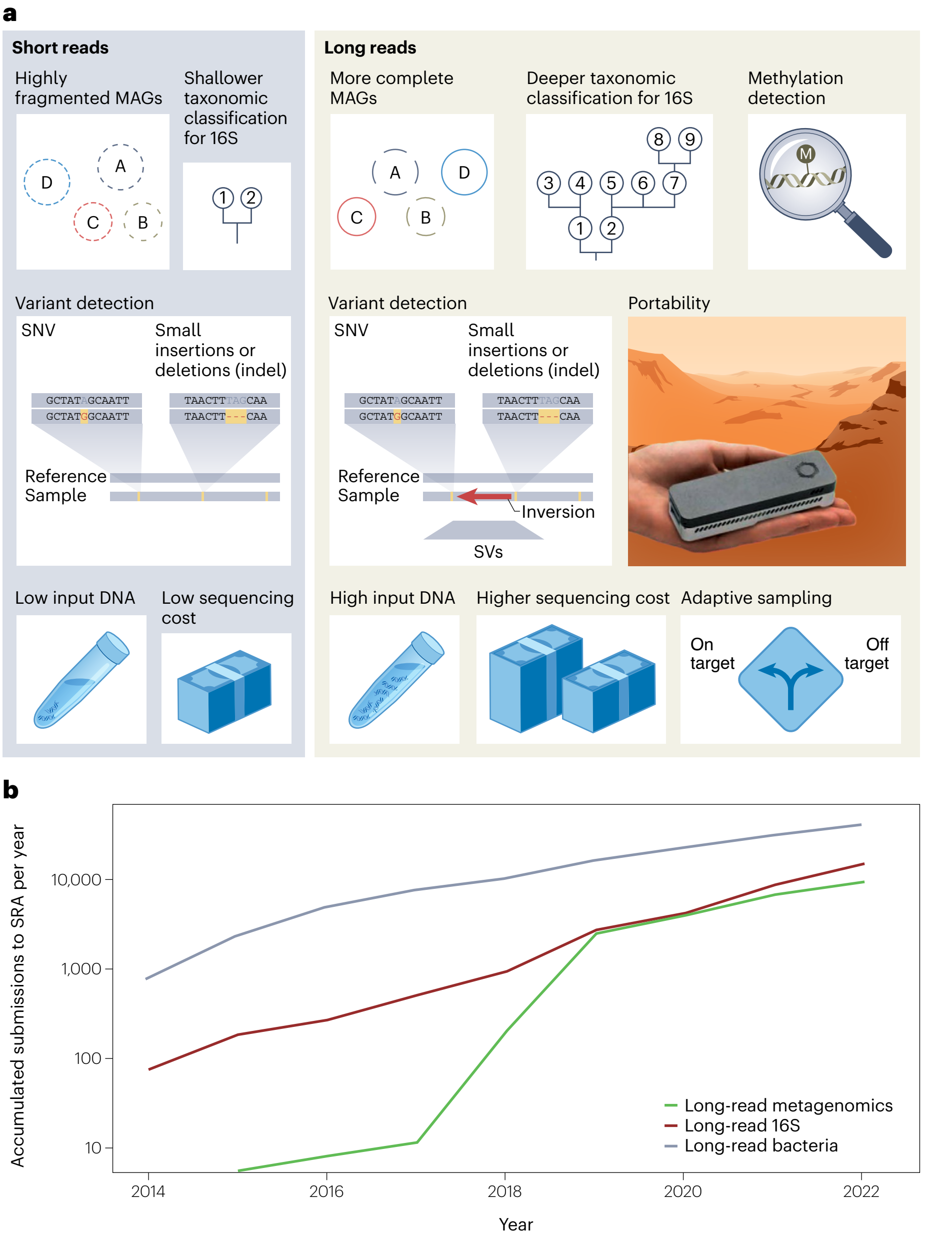

Fig. 1 | Overview of long reads in metagenomics. a, Differences between short-read and long-read technologies. Long reads have some advantages over short-read technologies. They can generate less fragmented genome assemblies, lower-level (species/strain) taxonomic characterization, DNA/RNA methylation pattern identification, large SV detection and highly portable sequencers. In contrast, short-read technologies still present overall cheaper sequencing costs and lower DNA input requirements due to the amplification step in library preparation. Image credit: Oxford Nanopore Technologies plc. **b**, The growth of long-read-related submission to the Sequence Read Archive (SRA) in recent years. Long-read platforms (ONT and PacBio SMRT) are being more widely used each year. The plot represents the accumulated number of data submissions related to each tag to the SRA each year.

the sequencing results. New preparation methods are constantly being developed together with low-input kits[12,20].

Contamination mitigation and targeted approaches in long-read metagenomics

Contamination from the laboratory apparatus or reagent kits can dramatically impact the result of metagenomic studies[21]. In clinical samples, contamination with host DNA is another important issue that might lower the detection of pathogens. There are currently many methods for dealing with contamination both before and after sequencing[20]. There are also commercial kits available to extract enriched microbial DNA based on differential methylation patterns between the host and microbiome. Bioinformatic tools have been developed for detecting and removing contamination or host reads from the sample. This is needed to avoid wrongful prediction of, for example, variants. Methods such as Decontam[22] and Recentrifuge[23] use comparisons between samples and negative controls to identify and remove contaminants from sequencing data.

Another solution is using a targeted instead of an untargeted approach. The latter is more exploratory as one captures the entire set of organisms but might suffer from higher requirements of coverage and sample input. Thus, it is common to pursue a targeted approach where scientists are interested in a defined set of organisms[19]. A common strategy for targeting is to amplify the whole genetic material in the sample through PCR techniques, such as multiple displacement amplification[10]. These methods, however, can generate a new problem, as they can create a high number of chimeric reads that are difficult to interpret. Some bioinformatic tools have been created to tackle this issue and have had moderate success[24]. Interestingly, the usage of a hybrid approach (combining long reads and short reads) helped circumvent this issue with chimeric long reads[10].

Alternatively, other methods for targeting organisms in a sample have been developed, such as capture panels[25], LAMP-seq[26] and adaptive sampling. The latter can be performed only by ONT, and it can be used to either deplete host reads or enrich for some microbial taxonomic groups during sequencing. This targeted enrichment/depletion is possible due to the 'ReadUntil' technology that allows the pores in the flow cells to selectively sequence DNA molecules based on genomes of interest[27]. As the DNA/RNA molecule is sequenced, the emerging read is aligned to genomes or genomic regions input by the user, and the software can decide whether to reject or accept it based on the parameters defined by the user (Fig. 1a).

Base calling and quality control

Base calling for PacBio is encapsulated into the sequencing software and Lima can be used for demultiplexing (Supplementary Table 1). For ONT, Guppy has replaced Bonito as the official base-calling tool, and a recently introduced and much faster tool called Dorado is likely to replace Guppy in the short future[28]. Currently, the ONT base callers can operate using three distinct algorithms chosen when running the software: FAST, high-accuracy and super-accuracy base calling. In that order, they provide increasing accuracy at the expense of speed and computational resources, with the speed of the base caller being a factor of the number of parameters in the neural network model[28]. Often after base calling, the read quality is estimated and some read filtering is performed, with failed reads not being reported. This filtering is based on different thresholds set during base calling, such as sample accuracy estimates (Q scores) or the number of subreads obtained from a DNA fragment in PacBio.

There are many software tools available that are used to obtain some important basic statistics about the sequencing run, such as read length distribution, the presence of adaptor sequence in the reads or contamination with small fragments. Read filtering also happens at this stage, and it is normally performed to remove low-quality reads, adaptors and reads that are too short. Ultimately, the scientist analyzing these data requires an understanding of potential biases of the data, which informs the potential limitations of the analysis and potential conclusions. Some of the tools designed for short reads are compatible with long-read technology. For instance, FastQC also works with ONT and PacBio long reads, and can provide a range of quality metrics for each read. Nanoplot[29] is a tool from the Nanopack package specific for ONT sequencing data, and can also provide quality-control statistics, metrics, processing and visualization for long reads. Guppy also provides simple quality-control reports about read sizes and other full metrics. Similarly, the Lima demultiplexing software from PacBio detects and trims adaptor sequences from PacBio reads. All tools are reviewed in Supplementary Table 1. These tools can be used either separately or in combination to perform quality control of long-read sequencing data. The importance of these methods cannot be overstated and is often crucial for the later success of an experiment.

Analysis using long-read sequencing

Given the unique aspects of long reads, new methods have been developed to enhance metagenomic analysis; selecting which methods to use is an integral part of the experimental design from the start (Fig. 2). Selecting appropriate tools for analysis and assessing the quality of the results is critical, and it is recommended to refer to benchmark studies or community-driven initiatives like CAMI/CAMI2 (ref. 30) to guide the choice of computational methods.

Depending on the experimental objectives, metagenomic approaches vary in their scope. Traditionally, when conducting taxonomic characterization, the emphasis was placed on analyzing marker genes, such as 16S/18S rRNA. However, contemporary approaches have broadened their focus to encompass whole-genome sequencing and, consequently, better discrimination of sub-strain-level genetic variations. Another strategy involves assembling individual genomes within metagenomic samples, requiring specialized methods to resolve ortholog clusters and correct potential analysis errors. Additionally, mapping reads to known references offers the advantage of characterizing low-frequency variants within the samples. In the following sections, we will provide an overview of these approaches and essential insights.

Taxonomic profiling with 16S/18S/ITS

Taxonomic characterization and quantification within a sample can be accomplished through two primary methodologies in metagenomics: 16S/18S/ITS rRNA sequencing (commonly known as marker gene or amplicon sequencing) and metagenomic shotgun whole-genome sequencing. In the 16S/18S/ITS approach, DNA is selectively amplified using PCR from specific marker genes, including the 16S rRNA gene for prokaryotes, the 18S rRNA gene for eukaryotes and the ITS regions for fungal identification. The resulting amplicons are then sequenced and aligned to reference databases, such as the NCBI 16S database[31], SILVA[32], RDP[33] and Greengenes[34], which represent diverse taxonomic groups. These marker genes serve as molecular signatures, enabling researchers to accurately identify, classify and quantify microorganisms, providing valuable insights into the taxonomic structure of complex microbial communities. Generally, these databases are not complete, and classification of organisms that are not present there will fail. While this approach has many limitations compared to whole-genome approaches, such as being more susceptible to primer biases, lower sensitivity to shallower taxonomic classification and unsuitability for organisms without marker genes (for example, viruses), it is still being often used due to its efficiency and sensitiveness in bacterial taxonomic profiling in complex microbiomes and its low price advantage[35]. To the extent of our knowledge, there are two computational tools designed for long reads to perform 16S taxonomic profiling—Emu[36] and NanoClust[37]—although most short-read methods can also work with long reads.

Previous comparisons between short-read and long-read 16S/18S/ITS analysis showed long reads achieve improved performance at the genus and species levels[9]. The higher resolution offered by long reads enhances the 16S rRNA classification and reduces the number of reads that cannot be classified or cannot be ascribed to lower-level taxa[38]. More importantly, long reads have the potential to uncover complete 16S rRNA sequences from microbial 'dark matter' with a higher resolution and reliability of classification[38].

Taxonomic characterization and abundance profiling in the whole-genome shotgun metagenomic approach

Whole-genome shotgun metagenomic profiling encompasses the sequencing of the complete genetic material within a mixed microbial sample, bypassing the need for prior selection or amplification of specific marker genes. Much like the 16S/18S/ITS approach, sequencing reads are matched against comprehensive whole-genome databases, such as IMG/M[39], MG-RAST[40] and NCBI's RefSeq[31]. Based on this matching, these reads are subsequently categorized into taxonomic groups, profiled and abundance can be evaluated.

There are multiple reference-based metagenomic tools that were designed specifically for long reads, such as Metamaps[41], Megan-LR[42], MMseqs2 (ref. 43), CDKAM[44] and BugSeq[45]. These reference-based profilers utilize machine learning or statistical models to assign taxonomic labels to genetic sequences, differently from genome aligners, which use algorithmic approaches such as dynamic programming or seed-and-extend methods to find optimal alignments or similarities between sequences. MetaMaps uses a reference-based approach,

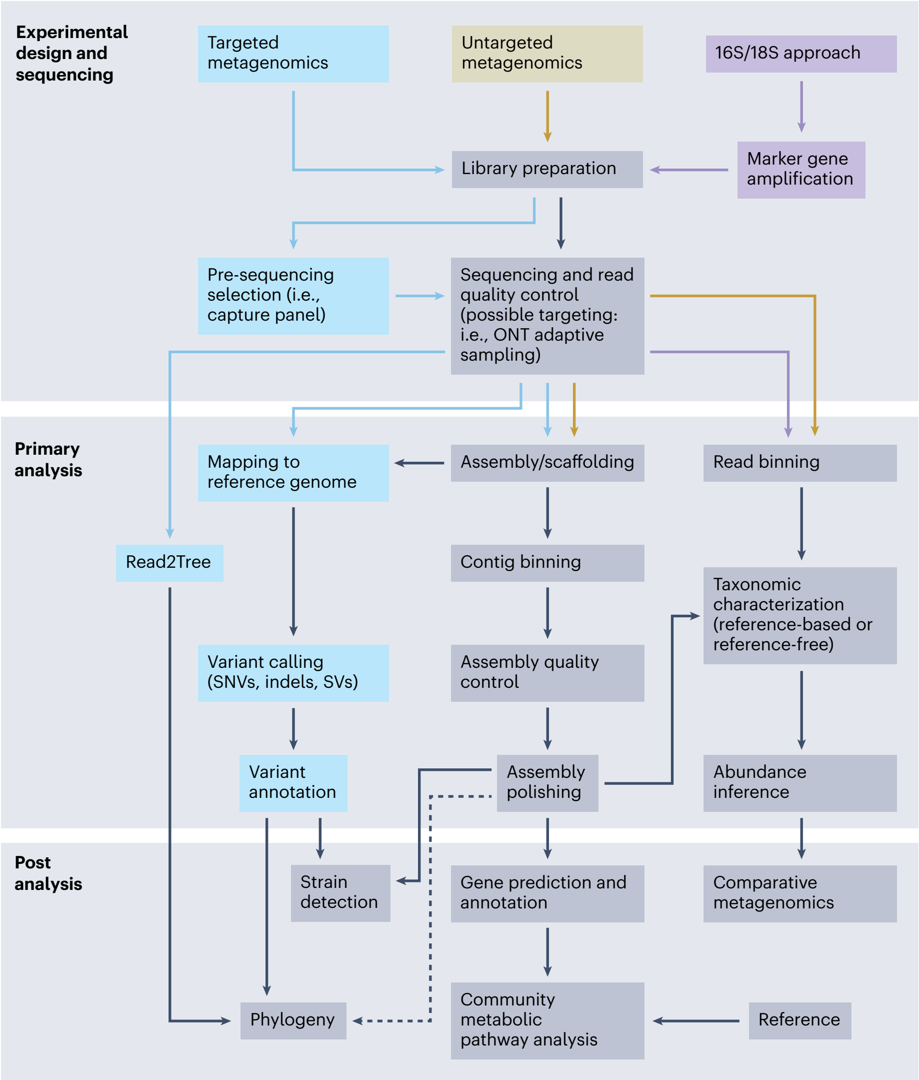
Experimental design and sequencing
Targeted metagenomics
Untargeted metagenomics
16S/18S approach
Library preparation
Marker gene amplification
Pre-sequencing selection (i.e., capture panel)
Sequencing and read quality control (possible targeting: i.e., ONT adaptive sampling)
Primary analysis
Mapping to reference genome
Assembly/scaffolding
Read binning
Read2Tree
Contig binning
Taxonomic characterization (reference-based or reference-free)
Variant calling (SNVs, indels, SVs)
Assembly quality control
Variant annotation
Assembly polishing
Abundance inference
Post analysis
Strain detection
Gene prediction and annotation
Comparative metagenomics
Phylogeny
Community metabolic pathway analysis
Reference

Fig. 2 | A generalized decision tree for metagenomic studies. When embarking on a metagenomic study, one of the first decisions a researcher must make is whether to use a targeted or an untargeted approach. A targeted approach involves sequencing a specific organism or group of organisms, often requiring enriching the target organism's genetic material from the microbiome sample using tiling amplicon panels, adaptive sampling or capture panels and probe designs. In contrast, an untargeted approach involves sequencing the entire population without prior selection. Another option to untargeted metagenomics is the 16S/18S rRNA approach, which involves sequencing the amplicons of a set of marker genes using primers specific to conserved regions of these genes to sequence the maximum number of organisms possible. The choice of approach considerably affects the available analyses and hypotheses that can be tested in each experiment. In this Review, we describe three main metagenomic analysis pipelines: mapping, de novo assembly, and taxonomic characterization. Each pipeline is more appropriate for different studies based on their objectives. Finally, we discuss post-sequencing steps that can be taken based on the different designs proposed. Dashed line arrows represent indirect processes that require other intermediate steps.

aligning sequences to a database, together with reference-free methods such as nucleotide composition to identify taxonomic groups. MEGAN-LR uses the alignments performed by other software to assign reads to a taxonomic group using a lowest common ancestor (LCA) algorithm. It can use a nucleotide sequence alignment, such as those produced by Minimap2 (ref. 46) or NGMLR[4], or a protein sequence aligned from the translation of the reads, such as those obtained from DIAMOND[47]. MMseqs2 works by extracting all protein fragments in six frames, filtering them, aligning them to a reference protein database, and using an LCA algorithm to assign the reads to specific taxa. CDKAM uses inexact *k*-mer matches to compensate for the increased error rate of long reads and identify matches in a reference database. BugSeq is a cloud platform and presents two different internal pipelines (V1 and V2), with one of those being auto-selected during the run. Both use Minimap2 alignments to a database, but in V1 it is followed by Bayesian reassignment and LCA identification, whereas in V2 the alignment is followed by LCA identification and abundance calculation[19] (Supplementary Table 1).

A recent benchmark performed using long reads and short reads, and with most of these tools, concluded that: (i) long reads increased precision of the calls; (ii) short-read methods used with long-read data resulted in high rates of false positives, specially Kraken2 and Centrifuge; (iii) most tools presented low precision, which could be increased with filtering, but at the cost of reducing recall; and (iv) the best-performing tools for long reads were Sourmash, BugSeq and MEGAN-LR (using either Minimap2 or DIAMOND for alignments), with high precision and recall without the need of read filtering[19]. This trade-off between precision and recall has also been observed in the CAMI challenge[30]. CDKAM was not included in this benchmark study. More recently, a study compared alignment and classification tools, finding similar or better accuracy and less RAM requirements for aligners such as Minimap2, despite being slower. Tools using nucleotide databases outperformed those using protein databases, with read length and database completeness influencing classification accuracy across datasets[48].

It is worth mentioning that protein sequence-based methods rely on pruning heuristics to consider only a certain number of open reading frames (ORFs). The longer the reads, the higher the chances of it containing frameshift sequencing errors and additional ORFs, which makes the challenge of identifying all ORFs across the reads nontrivial, and potentially missing some ORFs with those methods[49]. On the other hand, these methods require the presence of multiple ORFs in the read to work well, with their accuracy declining as reads get shorter[50]. This can be somewhat circumvented by fine-tuning of the parameters, which often involves a trade-off between sensitivity and execution time, with higher method sensitivity resulting in increased computational demands and processing time, without the guarantee of successful results[50].

De novo long-read assembly of genomes from metagenomes

The goal of metagenomic assembly methods is to reconstruct complete genomes for each microorganism present in a sample (called metagenome-assembled genomes or MAGs), thus avoiding potential representation or contamination issues with previously published reference genomes[51]. Bioinformatic tools perform this by tiling dovetailed reads that come from the same genomic region, forming contigs[5]. Likewise, contigs in close proximity inferred over reads can be tiled together to form longer stretches of sequences with unresolved regions, a process often referred to as scaffolding, until potentially forming chromosomes or whole genomes[5]. The most popular approaches for assemblies using long reads are to use the overlap-layout-consensus (OLC) method or de Bruijn graph (DBG) approach[5]. Within OLC, long reads are first aligned with each other to identify overlapping regions that are then joined together to form contigs. The DBG approach reduces this complexity to a *k*-mer space and guides the joining of reads in this way.

Repetitive regions and intergenomic repeats can impact on the efficiency of assembly tools, especially if these regions are longer than the overlapping length of reads or contigs[52,53]. Short reads can be used to detect variants between strains, but fail (due to the length) to relate (that is, phase) these variants into continuous haplotypes[53], which makes short-read assemblies notorious for contig fragmentation[10], and highly inefficient at assembling multi-copy DNA sequences, such as the 16S gene, mobile genetic elements (that is, transposons) and plasmids[54] (Fig. 1). Long-read methods can overcome this issue, but higher error rates present in some of the platforms could cause different problems. Initially, the alternative to minimize this effect was to perform hybrid assembly using long reads and short reads, but recent advancements to decrease long-read error rates have been making this approach obsolete[55].

The use of a hybrid approach to genome assembly. Hybrid assembly benefits from the advantages of long reads, and from the lower base-calling error rates from the short reads[56,57]. The addition of long reads to short-read-only assemblies generates an improvement in overall assembly statistics, contig size, binning and gene completeness[9,38], and increases the discovery of new species[38]. While hybrid assemblies show this remarkable advantage in comparison to short reads only, it comes with an increase in costs and necessary depth of sequencing. Besides, while long reads are efficient in dealing with repeats (intragenomic or intergenomic), short reads are not, and this can create ambiguities during the assembly process. Working with multiple platforms at the same time, can also introduce biases during assembly[5]. For instance, short reads can reintroduce G+C or primer biases from their different library preparation. Likewise, with the improvement in long-read technologies, error rates and costs per base have been dropping, allowing for increased sequencing coverage that can mitigate the long-read error rates. Besides, bioinformatics methods have evolved to circumvent these higher error rates using multiple strategies, either before or after assembly. Finally, hybrid assembly may not be beneficial when the accuracy of long reads is equivalent to that of short reads[9,58].

Some of the most popular tools for hybrid assembly are DBG2OLC[56], OPERA-MS[59] and Unicycler[57]. DBG2OLC converts a DBG representation of sequencing data into an OLC assembly by resolving ambiguities in the graph using long reads. OPERA-MS uses coverage-based clustering and Bayesian information criterion of clusters to perform short-read assembly, followed by overlaying with long reads to produce an assembly graph. Unicycler utilizes a combination of de novo assembly, read mapping and iterative correction steps to achieve accurate and complete reconstructions of bacterial genomes (Supplementary Table 1).

Long-read-based assembly methods for metagenomics. Due to improvements in sequencing technologies and decreased error rates, metagenome assemblies using long reads only have become sufficient to generate comprehensive and relatively error-free metagenome assemblies[55]. Figure 3 depicts a graph-based representation of a human gut metagenome assembly using exclusively PacBio HiFi reads. The graph distinguishes individual contigs by assigning each one a unique color. The presence of circularized contigs in the graph indicates the completeness of a genome sequence. This study was able to circularize 56 genomes from the human gut microbiome, highlighting the advancements made possible by novel long-read sequencing technologies and improved bioinformatics tools.

Nevertheless, due to error rates in the long reads (up to 1–3%; Table 1), multiple approaches rely on a read error correction before starting the actual assembly. Error correction can be important for downstream analysis and interpretation of metagenomic data. Correcting errors can increase the sensitivity and specificity of taxonomic classification and functional annotation and improve the accuracy of genome assembly and gene prediction. Furthermore, error correction can reduce the amount of noise in the data, which can lead to more accurate identification of rare taxa and the discovery of novel microorganisms[60]. However, error correction tools should be considered carefully before use in metagenomic studies, as although they can mitigate sequencing errors, they can also erase low-frequency variations and lower-frequency strains, affecting strain detection[61].

Read error correction is often performed by comparing the shorter reads to the longer reads in order to improve the quality of the longer reads (for example, Q score). Some assembler software provides built-in error correction features, such as MetaFlye[62], which uses a two-step overlap-based and assembly-based correction approach, and Canu[63], which uses a multiple sequence alignment approach. VeChat[60] uses a different approach—a variation graphs method to perform haplotype-aware error correction of reads.

The main tools designed for metagenome assembly using long reads are hifiasm-meta[64], Canu[63] and MetaFlye[62] and pipelines such as Lathe[65]. Other long-read assembly tools not specifically designed for a metagenomic approach are also used, although a recent benchmarking study using metagenomic data showed a subpar performance from those compared to metagenomic-specific tools[61]. A benchmark between Canu and MetaFlye reported a more complete assembly using Canu[10], while others showed a slight advantage to MetaFlye[61], suggesting that these tools have similar efficacy, probably subject to differences in the samples. Hifiasm-meta[64] is an assembler that takes advantage of reduced error rates from recent long-read sequencers and consists of several steps; optional read selection, sequencing error correction, read overlapping, string graph construction and graph cleaning (Fig. 3). Using HiFi reads, hifiasm-meta produced more total MAGs and more complete, single-contig circular MAGs than Canu or metaFlye[64].

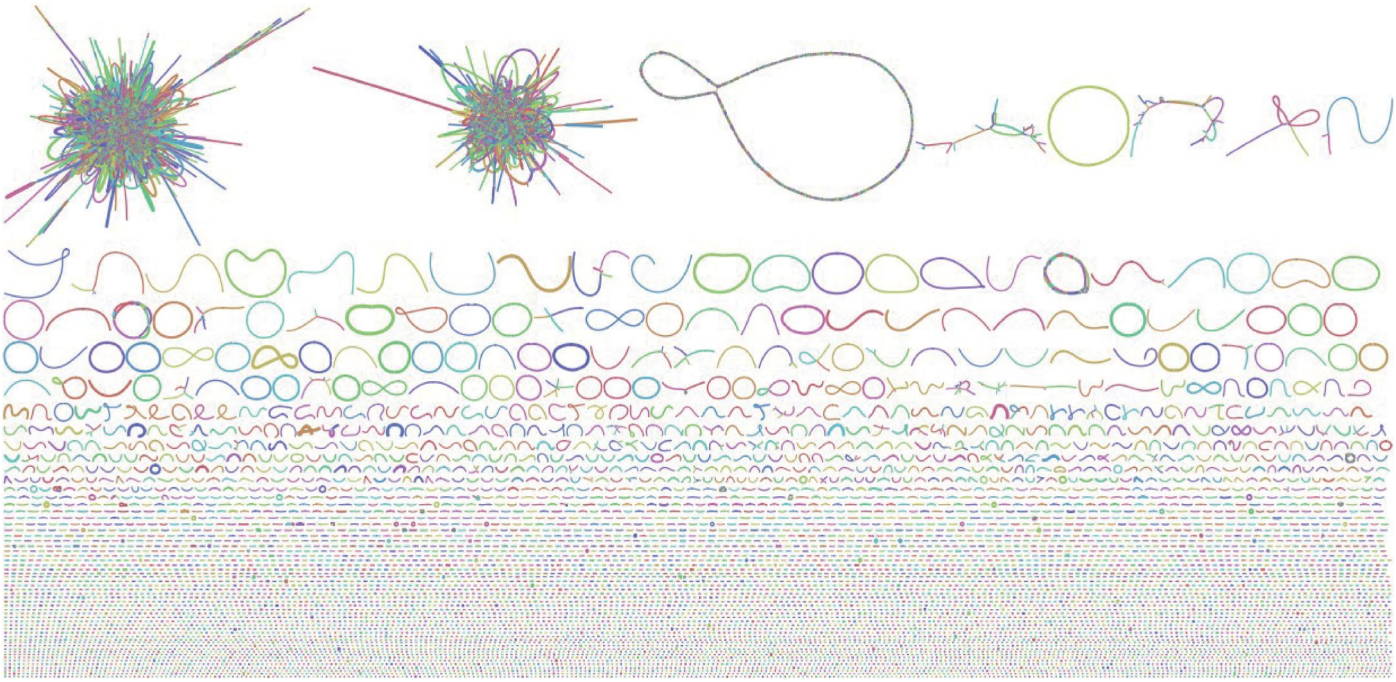

Fig. 3 | Graph representation of a metagenome assembly. Data were generated using HiFi reads and hifiasm-meta for long-read-only assembly. The figures represent all the contigs arranged by decreasing length, with each color representing a single contig. The sample is a commercially available pooled human gut reference (ZymoBIOMICS D6323). The dataset was generated with four SMRT Cells (8 M) on the PacBio Sequel IIe system, which yielded 11.9 million HiFi reads and 88.3 Gb of total data. There are 56 large circular contigs visible in the graph, ranging from 1.5 Mb to 6 Mb in size, along with numerous circular plasmids.

Once contigs are assembled, they can be investigated using read information to look for connections between each other, and then further merged to form bigger structures, called scaffolds[66]. A few assemblers such as MetaFlye have this feature already built in, but this can also be performed by scaffolding software such as MetaCarvel[67]. Some pipelines using combinations of different tools have also proven efficient; for instance, MetaBooster/MetaBooster-HiFi[68] combines error correction from VeChat with Canu metagenome assembly. Trycycler[69] is a long-read consensus tool designed specifically for bacterial genomes and constructs a consensus assembly based on multiple assemblies created by different tools. Other methods allow for post-assembly refinements. Strainberry[70] is a pipeline that takes a long-read metagenome assembly and performs variant calling and haplotype phasing to perform strain separation. stRainy[53] is a tool that can either perform assembly itself or take an already assembled set of MAGs as input and perform strain separation as well. These tools can be reviewed in Supplementary Table 1.

Taxonomic separation of MAGs (contig binning). Once the assembly is performed, the next step is the binning of contigs and scaffolds. Contig binning involves clustering of contigs into different bins based on their taxonomic or functional characteristics. The purpose of binning is to assign the contigs to the microbial organisms that they originated from, allowing for downstream analyses such as functional annotation and taxonomic profiling. Not all contigs need to be binned, particularly if they are already circularized and assigned to a specific organism or strain. However, it is generally recommended to bin as many contigs as possible, as this can improve the accuracy and completeness of downstream analyses[71]. It is worth noting that in multi-sample metagenomic datasets, using multi-coverage binning leads to higher-quality bins with reduced contamination compared to single-coverage binning. Direct comparison of both approaches using the same set of samples shows that multi-coverage binning outperforms single-coverage binning. It successfully identifies contaminant contigs and chimeric bins that other methods fail to detect[72].

Reference-free methods are generally recommended after assembly because reference-based tools will struggle with MAGs that are not present in their database. Tools designed for long-read contig binning already exist, such as MetaBCC-LR[73] and LRBinner[74]. Interestingly, these tools can also be used to perform binning of long reads directly as well, without previous assembly. MetaBCC-LR uses contig composition and coverage to infer the number of bins, and then a maximum likelihood framework to separate them into different bins. However, MetaBCC-LR uses a sampling strategy with large datasets, which can hinder the identification of low-frequency organisms. LRBinner combines compositional analysis with contig coverage (calculated by read alignment to the contigs) using Deep Learning and an iterative medoid clustering algorithm in addition to a distance-histogram-based clustering algorithm to separate the contigs into different bins without sub-sampling. Besides those, GraphMB[75] is a binner software designed for MAGs assembled from long reads and uses graph neural networks to incorporate the assembly graph into the binning process. Binnacle[76] uses scaffolding information to help separate genomes into bins. Some contig binners used for short-read assemblies will also work for this step, even if the assembly was produced with long reads. Short-read binner software has been exhaustively reviewed and benchmarked elsewhere[71] and are beyond the scope of this Review.

Binning in metagenomic samples can be challenging due to various factors such as horizontal gene transfers of DNA fragments like plasmids, as well as the presence of phages, which exhibit high sequence diversity and can exist as integrated proviruses within their host's DNA. To aid in the separation of these different sequences during contig binning, the identification of DNA methylation patterns using long-read sequencing data has been proposed[77]. Methylation pattern identification can also help in detecting chimeric reads[78], which could potentially improve contig binning. Furthermore, the genomic three-dimensional structure information provided by Hi-C can be utilized to enhance binning accuracy by leveraging proximity information, while also associating mobile elements with their respective host genomes[55].

Assembly polishing and quality control. After long-read MAG assembly, a common step for ONT is to 'polish the assembly', that is, to improve the accuracy of the draft assembly by using read data to correct errors in the assembly. Using short-read data is also possible, with polisher tools designed for that purpose. For technologies with low error rates, polishing with short reads does not provide improvements[55]. Besides, that involves another round of sequencing using a different platform, which can increase the costs. Hence, new tools were developed to use the same reads used to produce the assembly to perform polishing (Supplementary Table 1). A recent benchmark[79] study comparing these tools using *Escherichia coli* single-species sequencing data reached two main conclusions. First, ONT-only data can achieve the same quality of assembly without complementary short-read sequencing. Second, the most efficient tool for assembly polishing for single-species data was the reference-based HomoPolish[80], since a reference genome was already available. For unknown, multiple-species samples, their best results were with a combination of PEPPER[81] and Medaka (Supplementary Table 1).

Upon finishing the assembly, its quality can be measured using some specific metrics, such as N50, L50, genome completeness, genome size, assembly accuracy and contamination rate. N50 is the length of the shortest contig that, together with all the contigs of the assembly that have the same length or longer than it, covers 50% of the length of the genome assembled[5]. It is a measure of the contiguity of the assembly, and higher N50 values indicate longer contigs. L50 is the number of contigs needed to cover 50% of the genome assembly. It is also a measure of contiguity, and lower L50 values indicate fewer but longer contigs. While these metrics provide valuable insights into assembly length, they do not inherently guarantee its actual quality or completeness[5]. One way to assess this is to measure the number of unique single-copy genes (SCGs) that should be present. Contamination or assembly errors can be identified by the lack or multiple number of SCG copies[82]. Of course, caution is needed when assessing samples with multiple species, where only the lack of SCGs can be assessed[5].

These metrics can be obtained by a variety of tools, and many of those are also reviewed elsewhere[11,71]. BUSCO[83] uses a database of ortholog genomes (OrthoDB) to estimate complexity and redundancy of assembled genomes. It generates reports containing meaningful metrics that complement other statistics related to contig contiguity. Inspector[84] is a long-read de novo assembly evaluator that uses consensus sequences derived from raw reads covering erroneous regions in a reference-free way. It reports generic metrics and can accurately identify both large-scale and small-scale assembly errors. CheckM2 (ref. 85) provides robust estimates of genome completeness by using co-located sets of genes that are ubiquitous and SCGs within a phylogenetic lineage. They can also assess contamination, which is derived from the number(s) of SCGs present in the genome. DeepMAsED[86] uses a deep learning approach to detect misassembled contigs in a reference-free way. Taxonomic characterization tools such as Kraken2 (ref. 87) can also be used to identify and filter MAGs that match or don't match a taxonomic cluster of interest. Merqury[88] is a reference-free tool that works by comparing *k*-mers in an assembly to those found in unassembled high-accuracy reads and then estimating base-level accuracy and completeness (Supplementary Table 1).

Bin taxonomic classification. After the metagenomes are assembled and binned, the next challenge is the quantification and correct classification of organisms in a mixed metagenomic sample. The Genome Taxonomy Database Toolkit (GTDB-Tk)[89] is the gold-standard method to assign binned MAGs to specific taxonomic groups. MetaPhlAn4 (ref. 90) now incorporates MAG assembly followed by taxonomic characterization using the MetaPhlAn database. Abundance information can be inferred using tools like MEGAN-LR[42]. Novel organisms, however, might not be present in these databases. In that case, there are tools that can be used for the detection and annotation of ORFs. Some of the tools work directly on the reads, such as MMseqs2 (ref. 43) and MEGAN-LR[42] (together with DIAMOND[47]), while others can detect ORFs in MAGs, such as Prokka[91], PGAP[92] and SeqScreen-Nano[50]. CAT/BAT tools set is a pipeline for taxonomic classification of contigs and MAGs (bins), involving gene calling, ORF mapping and voting-based classification of contigs/MAGs, applicable to both known and unknown microorganisms in a sample[93]. Gene prediction and annotation have also been reviewed elsewhere[71] and are not specific to long reads. One special note is that these classification methods rely on databases that can sometimes be misleading due to contaminants or other biases from older assemblies[51]. Clearly, improvements are needed to avoid such biases with more and more novel species being discovered.

Reference-based analysis approaches of metagenomic samples

A mapping-based approach compares the raw reads directly to a reference genome, ideally taken from the same species or a very close relative. Its limitation is that one relies on reference genomes that can be unavailable or are only partially resolved. Nevertheless, mapping-based approaches have multiple advantages as they allow the identification of low-frequency mutations, easier comparison across multiple samples from the same, for example, pathogen and fewer constraints on coverage or read length[94]. Consequently, sequences absent in the reference, such as plasmids, might be missed in the analysis. A way to circumvent this is to perform this analysis in combination with assembly methods, creating a sample's reference genome[95].

A mapping approach is most commonly used with a targeted metagenomic approach (Fig. 2), where read enrichment from a specific organism or group of organisms is performed in the sample. One example is the detection of COVID-19 virus in patient samples[96] or wastewater[17]. Detection of new genomic variants is also important for pathogen surveillance, while tracking low-frequency intra-host variants provides important insights to elucidate host–virus population dynamics and transmission[97].

The mapping approach normally consists of aligning reads to a reference genome and then using these alignment files to quantify species presence or detect variations. There are multiple benefits from aligning long reads, such as higher alignment rates or overcoming ambiguity due to repeats or high homologous regions compared to short-read methods. In addition, structural variants are often easier to identify and resolve[5,7]. Similarly to assembly methods, one should also choose a long-read-alignment method carefully[5,7]. Long-read aligners often use a seed-and-chain paradigm, where multiple anchors are gathered and chained together to form a candidate extending procedure, allowing often for more accurate mapping than short reads. Additionally, long-read aligners have incorporated sketching techniques, borrowed from comparative genomics to improve throughput and efficiency in handling the large number of reads[98]. The most widespread aligners used are Minimap2 (ref. 46) and NGMLR[4]. While most of these methods are developed outside metagenomics, they can be successfully used in this field[5].

Variant calling methods use alignments to identify locations where the sequence from the sample differs from the reference sequence. They use statistical algorithms to distinguish true variants from sequencing errors and other biases. The variants are often classified into single nucleotide variants (SNVs), small insertions/deletions (indels, typically <50 bp) or larger structural variants (SVs)[5,7]. Clair3 (ref. 99) uses a pileup and full-alignment, two-module method to detect SNVs and indels. DeepVariant[100] uses a deep convolutional neural network to call SNVs and indels in alignment files. Medaka uses a local realignment approach to detect and genotype variants and then applies a neural network to estimate allele frequency. It is especially designed for SNVs and indels, but it also works on SVs. NanoCaller[101] is a deep learning method that detects SNVs using long-range haplotype information, then phases long reads with called SNVs and calls indels with local realignment. Lofreq[102] is designed to work on any kind of sequencing technology and models sequencing run-specific error rates to accurately call SNVs and indels. Variabel[97] is a variant call filtering tool to be used after variant calling and designed for viral samples, improving the prediction of low-frequency variants in ONT data (Supplementary Table 1). Variants are then annotated to determine their impact on gene structure and protein sequence. The most used variant annotation tools for SNVs and indels are SNPEff[103], ANNOVAR[104] and Ensembl-VEP[105] (Supplementary Table 1).

Some tools have been created specifically for viral surveillance, and they can detect new variants of concern in samples, such as wastewater. Data from mixed strains in the sample can be deconvoluted to define the individual strains and their abundance, and phylogenetic distances can be measured to create a phylogenetic tree between strains. An example of such a tool is Read2tree[106], which as the name suggests, skips all these steps going straight from the raw reads to a phylogenetic tree, although it does perform alignments using Minimap2 to a database of genome-wide reference orthologous groups and other steps internally. When performing phylogenetic analysis, it should be taken into consideration that microorganisms such as bacteria can exchange genetic material via horizontal gene transfer, recombination and other mechanisms, which can result in complex and reticulate evolutionary relationships, and are better represented by phylogenetic networks rather than by traditional bifurcating phylogenetic trees[107].

SV analysis. SVs are ubiquitous in both individual bacteria and across microbial communities inhabiting human hosts[108]. SVs are often defined as >50-bp genomic alterations including insertions, inversions, deletions, duplications and translocations[66], and have shown a profound impact on eukaryotes (for example, human population diversity, diseases and other phenotypes). In a phenomenon specific to microbial genomes, bacterial genomes can undergo horizontal gene transfer, a process central to bacterial evolution and adaptation[3]. Besides, viral genomes exhibit complex transcriptional patterns and a high propensity for recombination, distinguishing them from other biological entities[109]. While there are numerous studies that have analyzed SVs in microbial genomes, to date only a few studies have analyzed SVs across metagenomic samples[108]. This occurs despite evidence of their occurrence and potential impacts on viruses, bacteria and others. Thus, most of the existing SV methods are designed with diploid genomes (for example, humans) in mind.

In general, there are two different tool categories: assembly-based and mapping-based methods. To name a few assembly-based SV calling methods, Dipcall[110] or Mummer[111] have demonstrated considerable reliability. Whereas for mapping-based methods, Sniffles2 (ref. 112) and SVIM-asm[113] are excellent examples. The former uses an SV scoring scheme to exclude false SVs, while incorporating the detection of low-frequency SVs across different datasets. The latter uses split-read and read-depth methods to identify SVs, and it can detect complex events. The interpretation of SVs in a metagenomics context is challenging but also presents a great opportunity for future studies. In eukaryotes, much was learned by assessing the SV frequency in specific populations, while similar metagenomic studies are nonexistent. SV annotation can be performed by Ensembl-VEP[105] and AnnotSV[114] (Supplementary Table 1). Comparing mapping-based SVs across metagenomic samples might be possible with SURVIVOR[115] and Truvari[116], but again more specialized methods are needed for metagenomics.

While SV detection with long reads in individual microbial genomes is straightforward, SV detecting in microbiomes containing a diverse set of microorganisms presents a substantial challenge due to unknown reference genomes and the presence of mixed microbial strains within the sample. A recent method in this space, Rhea[117], forgoes reference genomes and MAGs by building a single metagenome coassembly graph constructed from temporally sampled microbiomes. Rhea then maps the long reads to the coassembly graph to infer SVs between time points based on graph-based variant detection.

Utilization of epigenetic signals in metagenomic analysis

The role of DNA and RNA modifications in both prokaryotes and eukaryotes, as well as the available methods to detect them, have been thoroughly reviewed by Kong et al.[6]. The most common of DNA methylation-based modifications are N^6-methyladenine (6mA), N^4-methylcytosine (4mC), 5-hydroxymethylcytosine (5hmC) and 5-methylcytosine (5mC), but others exist. 5mC is the dominant modification in eukaryotes, while 6mA is the most prevalent in prokaryotes[6]. Both RNA and DNA[6] viruses can also present genomic methylation, with m^6A (N^6-methyladenosine RNA modification, as opposed to 6mA, which is a DNA modification) as an important marker in RNA viruses[118]. Both long-read platforms provide methylation information on CpGs for DNA sequencing, outperforming bisulfite-based, short-read sequencing for methylation detection by identifying methylation in diverse bases and obviating the need for a reference genome[119]. Nanopore sequencing technology allows for direct RNA-seq, making it possible to detect these RNA modifications steadily[118]. Besides the study of epigenetic modifications being important in itself, the detection of these modifications can facilitate taxonomic characterization, binning and strain separation during metagenomic studies[77]. Currently, the literature and available tools are mostly focused on the detection of DNA methylation in eukaryotes, and hence on 5mC modifications[6].

Tools tailored for metagenomics must be able to detect 5mC, 6mA and 4mC modifications, thus being able to detect both prokaryote and eukaryote microorganisms, such as fungi. Among them, Nanopolish[120] can detect several types of DNA modifications, using a statistical model to analyze the raw signal data generated by ONT. DeepSignal[121] is a deep learning-based software that can detect several types of DNA modifications, including 5mC and 6mA, using a convolutional neural network to analyze the raw signal ONT data. DeepMP[122] is a convolutional neural network-based model that takes information from ONT raw signals and base-calling errors to detect whether a given motif in a read is methylated, being able to detect 6mA and 5mC modifications. Remora (Supplementary Table 1) can identify 4mC, 5mC and 6mA modifications in ONT reads after base calling. Most of these tools depend on the identification of the genomic context to detect methylation, such as 5mC in CpG islands, or 6mA at GATC motifs. Nanodisco[123] is a tool to detect DNA methylation in prokaryotes regardless of the genomic context, and that has shown the three types of DNA methylation in diverse sequence contexts (Supplementary Table 1). It is important to note that most bioinformatic methods for long-read methylation detection typically require a negative control, such as synthetically amplified whole genomes, to ensure accurate and reliable methylation detection. For the detection of RNA methylation markers from direct RNA-seq, in particular the m^6A modification, there are few state-of-the-art tools. Of note, Nanocompore[124] and Epinano[125] are well-consolidated tools. While methylation analysis has the potential to provide deeper insights, it is not as commonly used by the community yet.

Challenges and future directions

In this review, we reported the state-of-the-art metagenomics methods utilizing long-read sequencing technologies. We discussed steps from experimental design across sequencing and analytical approaches and mentioned several secondary analysis approaches as well. Together with these individual steps and suggestions, we provide an extensive list of methodologies for the reader. Although we could not list all available methods designed for long reads, we highlighted some of the most relevant for each approach.

Metagenomic analysis faces a multitude of challenges. Foremost among them is the quest for accurate and unbiased species identification within a given sample. The effectiveness of this endeavor is inevitably tied to the chosen detection method, and often, different methods yield conflicting results. This might be improved with future developments where long reads hold a strong promise to disentangle this information. Furthermore, the absence of universally accepted benchmark samples in metagenomics, analogous to the well-established GIAB[66] for human genomes, poses a substantial hurdle. While mock samples exist, they frequently suffer from contamination and impurities, thereby complicating the enforcement of stringent standards for strain identifications. The persistent challenge of metagenomic sample classification lies in the ability to discern sub-strain variations, a common yet elusive target.

Some aspects of long-read sequencing remain underutilized, such as the simultaneous assessment of nucleotide and methylation information. Recent studies in eukaryotes showed that methylation can be utilized to distinguish even haplotypes[126]. This could easily be extended to improve metagenomic analysis to detect even below strain levels. Furthermore, in the eukaryotic world, long reads considerably improved the study of SVs, which led to multiple discoveries such as speciation events and other phenotypic impacts[7]. The current state of structural variation detection in metagenomic samples is unfortunately overlooked. The primary challenge lies in our inability to generate sufficiently high-quality MAG assemblies to comprehensively explore these phenomena within microbiomes. The current methods in use are mainly designed for diploid genomes (that is, human samples) and ignore challenges of cross-species mapping or other signals. Incorporating methylation and SV information into metagenomic algorithms promises to yield new biological insights and enhance the effectiveness of long-read technologies, thereby remarkably advancing metagenomic analysis.

Furthermore, the ongoing expansion of genomic databases presents a unique opportunity for advancing the precision of long-read taxonomic classification algorithms. As these databases grow, they encompass an increasingly diverse array of genomes from various microorganisms. Long-read sequencing, when coupled with multi-omics data integration, can harness this wealth of genomic information to enhance taxonomic classification accuracy. The availability of more comprehensive reference genomes, derived from long-read sequencing technologies, contributes to a better representation of the microbial world. With extensive genomic coverage and a broader range of genetic markers, the taxonomic resolution achievable by these algorithms is poised to improve substantially. This convergence of growing genomic databases, long-read sequencing and multi-omics integration underscores the potential for achieving unprecedented taxonomic precision in metagenomic analyses.

In recent years, there have been promising innovations in long-read sequencing technologies, marked by increased yield and reduced sample requirements. These innovations, however, are still on the path to fulfilling their full potential. Notably, breakthroughs like the telomere-to-telomere (T2T) assemblies[127], primarily applied within the realm of eukaryotes, hold the promise of translation into metagenomic practices. Undoubtedly, these advancements will make an important impact on the field, necessitating further computational methods.

In the clinical setting, the decreasing prices, runtime and portability of sequencers, together with the development of accessible bioinformatic pipelines, can make long-read sequencing ubiquitous in the physicians' toolbox, and a useful instrument in personalized medicine helping in the diagnosis of infections by sequencing patient samples, as well as help choosing the best treatment by identifying antibiotic resistance genes. In addition, targeted metagenomics has shown to be a useful asset for pathogen surveillance, as evidenced by using long-read sequencing on targeted metagenomic analysis of wastewater, which proved very efficacious and emerged as a great tool in improving viral surveillance in the microbial community[17].

In conclusion, long-read sequencing has considerably impacted the field of metagenomics and beyond, paving the way for groundbreaking research in various disciplines. While it continues to evolve, with new developments and advancements enhancing its capabilities, some challenges such as error rates, sample requirements and cost persist. Nevertheless, long-read sequencing has firmly established its position and is poised to revolutionize another frontier in life sciences with unwavering potential.

References

1. Edwards, R. A. et al. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57 (2006).
2. Tamburini, F. B. et al. Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. *Nat. Commun.* **13**, 926 (2022).
3. van Almsick, V., Schuler, F., Mellmann, A. & Schwierzeck, V. The use of long-read sequencing technologies in infection control: horizontal transfer of a $bla_{CTX\text{-}M\text{-}27}$ containing lncFII plasmid in a patient screening sample. *Microorganisms* **10**, 491 (2022).
4. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
5. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
6. Kong, Y., Mead, E. A. & Fang, G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat. Rev. Genet*. 10.1038/s41576-022-00559-5 (2023).
7. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
8. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
9. Gehrig, J. L. et al. Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microb. Genom.* **8**, 000794 (2022).
10. Kiguchi, Y., Nishijima, S., Kumar, N., Hattori, M. & Suda, W. Long-read metagenomics of multiple displacement amplified DNA of low-biomass human gut phageomes by SACRA pre-processing chimeric reads. *DNA Res.* **28**, dsab019 (2021).
11. Olson, N. D. et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* **20**, 1140–1150 (2019).
12. Ni, Y., Liu, X., Simeneh, Z. M., Yang, M. & Li, R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput. Struct. Biotechnol. J.* **21**, 2352–2364 (2023).
13. Castro-Wallace, S. L. et al. Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* **7**, 18022 (2017).
14. Cheng, H. et al. A rapid bacterial pathogen and antimicrobial resistance diagnosis workflow using Oxford nanopore adaptive sequencing method. *Brief. Bioinform.* **23**, bbac453 (2022).
15. Zhang, L. et al. Rapid detection of bacterial pathogens and antimicrobial resistance genes in clinical urine samples with urinary tract infection by metagenomic nanopore sequencing. *Front. Microbiol.* **13**, 858777 (2022).
16. Isidro, J. et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. *Nat. Med.* **28**, 1569–1572 (2022).
17. Karthikeyan, S. et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).

18. Gaulke, C. A. et al. Evaluation of the effects of library preparation procedure and sample characteristics on the accuracy of metagenomic profiles. *mSystems* **6**, e0044021 (2021).
19. Portik, D. M., Brown, C. T. & Pierce-Ward, N. T. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics* **23**, 541 (2022).
20. Wang, C. et al. Toward efficient and high-fidelity metagenomic data from sub-nanogram DNA: evaluation of library preparation and decontamination methods. *BMC Biol.* **20**, 225 (2022).
21. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
22. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
23. Martí, J. M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput. Biol.* **15**, e1006967 (2019).
24. Warris, S. et al. Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics* **19**, 798 (2018).
25. McCall, C. et al. Targeted metagenomic sequencing for detection of vertebrate viruses in wastewater for public health surveillance. *ACS EST Water* https://doi.org/10.1021/acsestwater.3c00183 (2023).
26. Ludwig, K. U. et al. LAMP-Seq enables sensitive, multiplexed COVID-19 diagnostics using molecular barcoding. *Nat. Biotechnol.* **39**, 1556–1562 (2021).
27. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754 (2016).
28. Samarakoon, H., Ferguson, J. M., Gamaarachchi, H. & Deveson, I. W. Accelerated nanopore basecalling with SLOW5 data format. *Bioinformatics* **39**, btad352 (2023).
29. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
30. Meyer, F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
31. Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
32. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
33. Cole, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
34. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
35. Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
36. Curry, K. D. et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nat. Methods* **19**, 845–853 (2022).
37. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* **37**, 1600–1601 (2021).
38. Zaragoza-Solas, A., Haro-Moreno, J. M., Rodriguez-Valera, F. & López-Pérez, M. Long-read metagenomics improves the recovery of viral diversity from complex natural marine samples. *mSystems* **7**, e0019222 (2022).
39. Chen, I.-M. A. et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
40. Keegan, K. P., Glass, E. M. & Meyer, F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol. Biol.* **1399**, 207–233 (2016).
41. Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* **10**, 3066 (2019).
42. Huson, D. H. et al. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol. Direct* **13**, 6 (2018).
43. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
44. Bui, V.-K. & Wei, C. CDKAM: a taxonomic classification tool using discriminative *k*-mers and approximate matching strategies. *BMC Bioinformatics* **21**, 468 (2020).
45. Fan, J., Huang, S. & Chorlton, S. D. BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics* **22**, 160 (2021).
46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
48. Marić, J., Križanović, K., Riondet, S., Nagarajan, N. & Šikić, M. Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *BMC Bioinformatics* **25**, 15 (2024).
49. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).
50. Balaji, A. et al. SeqScreen-Nano: a computational platform for rapid, in-field characterization of previously unseen pathogens. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.10.528096 (2023).
51. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**, 954–960 (2019).
52. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
53. Kazantseva, E., Donmez, A., Pop, M. & Kolmogorov, M. stRainy: assembly-based metagenomic strain phasing using long reads. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.31.526521 (2023).
54. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb. Genom.* **6**, mgen000436 (2020).
55. Bickhart, D. M. et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* **40**, 711–719 (2022).
56. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).
57. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
58. Haro-Moreno, J. M., López-Pérez, M. & Rodriguez-Valera, F. Enhanced recovery of microbial genes and genomes from a marine water column using long-read metagenomics. *Front. Microbiol.* **12**, 708782 (2021).
59. Bertrand, D. et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
60. Luo, X., Kang, X. & Schönhuth, A. VeChat: correcting errors in long reads using variation graphs. *Nat. Commun.* **13**, 6657 (2022).
61. Zhang, Z., Yang, C., Veldsman, W. P., Fang, X. & Zhang, L. Benchmarking genome assembly methods on metagenomic sequencing data. Brief. Bioinform. 24, (2023).

62. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
63. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
64. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat. Methods* **19**, 671–674 (2022).
65. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).
66. Majidian, S., Agustinho, D. P., Chin, C.-S., Sedlazeck, F. J. & Mahmoud, M. Genomic variant benchmark: if you cannot measure it, you cannot improve it. *Genome Biol.* **24**, 221 (2023).
67. Ghurye, J., Treangen, T., Fedarko, M., Hervey, W. J. 4th & Pop, M. MetaCarvel: linking assembly graph motifs to biological variants. *Genome Biol.* **20**, 174 (2019).
68. Luo, X., Kang, X. & Schönhuth, A. Enhancing long-read-based strain-aware metagenome assembly. *Front. Genet.* **13**, 868280 (2022).
69. Wick, R. R. et al. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol.* **22**, 266 (2021).
70. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat. Commun.* **12**, 4485 (2021).
71. Yang, C. et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
72. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
73. Wickramarachchi, A., Mallawaarachchi, V., Rajan, V. & Lin, Y. MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics* **36**, i3–i11 (2020).
74. Wickramarachchi, A. & Lin, Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol. Biol.* **17**, 14 (2022).
75. Lamurias, A., Sereika, M., Albertsen, M., Hose, K. & Nielsen, T. D. Metagenomic binning with assembly graph embeddings. *Bioinformatics* **38**, 4481–4487 (2022).
76. Muralidharan, H. S., Shah, N., Meisel, J. S. & Pop, M. Binnacle: using scaffolds to improve the contiguity and quality of metagenomic bins. *Front. Microbiol.* **12**, 638561 (2021).
77. Wilbanks, E. G. et al. Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J.* **16**, 1921–1931 (2022).
78. Berthelier, J. et al. Long-read direct RNA sequencing reveals epigenetic regulation of chimeric gene-transposon transcripts in *Arabidopsis thaliana*. *Nat. Commun.* **14**, 3248 (2023).
79. Lee, J. Y. et al. Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. *Sci. Rep.* **11**, 20740 (2021).
80. Huang, Y. -T., Liu, P. -Y. & Shih, P. -W. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. *Genome Biol.* **22**, 95 (2021).
81. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
82. Cornet, L. & Baurain, D. Contamination detection in genomic data: more is not enough. *Genome Biol.* **23**, 60 (2022).
83. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
84. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
85. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
86. Mineeva, O., Rojas-Carulla, M., Ley, R. E., Schölkopf, B. & Youngblut, N. D. DeepMAsED: evaluating the quality of metagenomic assemblies. *Bioinformatics* **36**, 3011–3017 (2020).
87. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
88. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
89. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
90. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. *Nat Biotechnol.* **41**, 1633–1644 (2023).
91. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
92. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
93. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
94. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).
95. Ajami, N. J., Wong, M. C., Ross, M. C., Lloyd, R. E. & Petrosino, J. F. Maximal viral information recovery from sequence data using VirMAP. *Nat. Commun.* **9**, 3205 (2018).
96. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921 (2020).
97. Liu, Y. et al. Rescuing low frequency variants within intra-host viral populations directly from Oxford Nanopore sequencing data. *Nat. Commun.* **13**, 1321 (2022).
98. Sahlin, K., Baudeau, T., Cazaux, B. & Marchet, C. A survey of mapping algorithms in the long-reads era. *Genome Biol.* **24**, 133 (2023).
99. Su, J., Zheng, Z., Ahmed, S. S., Lam, T.-W. & Luo, R. Clair3-trio: high-performance Nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. *Brief. Bioinform.* **23**, bbac301 (2022).
100. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
101. Ahsan, M. U., Liu, Q., Fang, L. & Wang, K. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biol.* **22**, 261 (2021).
102. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
103. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
104. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
105. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

106. Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J. & Dessimoz, C. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.* **42**, 139–147 (2024).
107. Corel, E. et al. Bipartite network analysis of gene sharings in the microbial world. *Mol. Biol. Evol.* **35**, 899–913 (2018).
108. Chen, L. et al. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).
109. Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F. & González-Candelas, F. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.* **30**, 296–307 (2015).
110. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
111. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
112. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02024-y (2024).
113. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
114. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
115. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
116. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
117. Curry, K. D. et al. Reference-free structural variant detection in microbiomes via long-read coassembly graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.25.577285 (2024).
118. Zhang, T. et al. N^6-methyladenosine RNA modification promotes viral genomic RNA stability and infection. *Nat. Commun.* **13**, 6576 (2022).
119. Barros-Silva, D., Joana Marques, C., Henrique, R. & Jerónimo, C. Profiling DNA methylation based on next-generation sequencing approaches: new insights and clinical applications. *Genes* **9**, 429 (2018).
120. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
121. Ni, P. et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **35**, 4586–4595 (2019).
122. Bonet, J. et al. DeepMP: a deep learning tool to detect DNA base modifications on Nanopore sequencing data. *Bioinformatics* **38**, 1235–1243 (2021).
123. Tourancheau, A., Mead, E. A., Zhang, X. -S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).
124. Leger, A. et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.* **12**, 7198 (2021).
125. Liu, H. et al. Accurate detection of m^{6}A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).
126. Fu, Y. et al. MethPhaser: methylation-based haplotype phasing of human genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.12.540573 (2023).
127. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
128. Pfeiffer, F. et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
129. Espinosa, E., Bautista, R., Larrosa, R. & Plata, O. Advancements in long-read genome sequencing technologies and algorithms. *Genomics* **116**, 110842 (2024).
130. Salamon, D. et al. Comparison of iSeq and MiSeq as the two platforms for 16S rRNA sequencing in the study of the gut of rat microbiome. *Appl. Microbiol. Biotechnol.* **106**, 7671–7681 (2022).
131. 41J Blog. Cost per gigabase. https://41j.com/blog/2022/09/cost-per-gigabase/ (2022).
132. Mastrorosa, F. K., Miller, D. E. & Eichler, E. E. Applications of long-read sequencing to Mendelian genetics. *Genome Med.* **15**, 42 (2023).

Acknowledgements

We thank D. M. Portik, for important feedback and kindly authorizing the use of Fig. 3. We also thank A. F. Pomerantz and K. D. Curry for their helpful feedback. This work was supported by the National Institute of Allergy and Infectious Diseases (grant nos. 1U19AI144297 and P01-AI152999) and by the National Science Foundation (grant nos. EF-2126387 and IIS-2239114). Y.F. and T.T. were supported in part by the National Institutes of Health NIAID award P01-AI152999. T.T. was also supported by the National Science Foundation grants EF-2126387, IIS-2239114 and CNS-1338099.

Competing interests

F.J.S. has received research funding from Illumina, PacBio, Genentech and Oxford Nanopore. V.K.M. is an employee of Genentech. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02262-1.

Correspondence and requests for materials should be addressed to Fritz J. Sedlazeck.

Peer review information *Nature Methods* thanks Ami Bhatt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.