nature reviews methods primers

https://doi.org/10.1038/s43586-024-00376-6

Primer

Check for updates

Analysis of metagenomic data

Shaopeng Liu[1,33], Judith S. Rodriguez[1,33], Viorel Munteanu[2,3,33], Cynthia Ronkowski[4], Nitesh Kumar Sharma[4], Mohammed Alser[4,5], Francesco Andreace[6,7], Ran Blekhman[8], Dagmara Błaszczyk[9], Rayan Chikhi[6], Keith A. Crandall[10], Katja Della Libera[8], Dallace Francis[11], Alina Frolova[12], Abigail Shahar Gancz[13,14], Naomi E. Huntley[14,15], Pooja Jaiswal[4], Tomasz Kosciolek[9], Pawel P. Łabaj[9], Wojciech Łabaj[16], Tu Luan[17,18], Christopher Mason[19,20,21,22], Ahmed M. Moustafa[23,24,25], Harihara Subrahmaniam Muralidharan[16], Onur Mutlu[26], Nika Mansouri Ghiasi[26], Ali Rahnavard[10], Fengzhu Sun[11], Shuchang Tian[14,27], Braden T. Tierney[19], Emily Van Syoc[14,15], Riccardo Vicedomini[6,7], Joseph P. Zackular[25,28,29], Alex Zelikovsky[3,5], Kinga Zielińska[9], Erika Ganda[14,30], Emily R. Davenport[1,14,30], Mihai Pop[19,31,34], David Koslicki[1,14,30,32,34] & Serghei Mangul[3,4,34]

Abstract

Metagenomics has revolutionized our understanding of microbial communities, offering unprecedented insights into their genetic and functional diversity across Earth's diverse ecosystems. Beyond their roles as environmental constituents, microbiomes act as symbionts, profoundly influencing the health and function of their host organisms. Given the inherent complexity of these communities and the diverse environments where they reside, the components of a metagenomics study must be carefully tailored to yield accurate results that are representative of the populations of interest. This Primer examines the methodological advancements and current practices that have shaped the field, from initial stages of sample collection and DNA extraction to the advanced bioinformatics tools employed for data analysis, with a particular focus on the profound impact of next-generation sequencing on the scale and accuracy of metagenomics studies. We critically assess the challenges and limitations inherent in metagenomics experimentation, available technologies and computational analysis methods. Beyond technical methodologies, we explore the application of metagenomics across various domains, including human health, agriculture and environmental monitoring. Looking ahead, we advocate for the development of more robust computational frameworks and enhanced interdisciplinary collaborations. This Primer serves as a comprehensive guide for advancing the precision and applicability of metagenomic studies, positioning them to address the complexities of microbial ecology and their broader implications for human health and environmental sustainability.

Sections

A full list of affiliations appears at the end of the paper. e-mail: serghei.mangul@gmail.com

Introduction

Metagenomics is an interdisciplinary field encompassing experimental and computational methods for analysing the genomic content and functional potential of microbial communities. Metagenomic studies typically begin with the collection of an environmental sample of interest, such as soil, water, blood or stool. After collection, the total DNA within the sample is extracted and sequenced using whole-genome shotgun sequencing to generate reads originating from random genomic loci, ultimately generating a metagenomic profile to obtain a more extensive understanding of the microorganisms in the sample. This approach contrasts with amplicon-based approaches for profiling bacteria that selectively amplify and sequence the 16S and 18S small subunit ribosomal RNA (SSU rRNA) marker genes (Supplementary Boxes 1 and 2).

Metagenomics originally emerged as a powerful tool for exploring the extensive microbial diversity within environmental samples such as soil[1] and water[2], providing unprecedented insights into the complexity and function of microbial communities in their natural habitats. As the field has evolved, metagenomics has become a foundational methodology, driving advances in microbial community research, and has been applied to critical areas such as human health, food safety, agriculture and biotechnology. Early metagenomic methods relied on Sanger sequencing, which involves the cloning of randomly fragmented DNA into culturable bacteria[3]. Current methods are based on next-generation sequencing (NGS), which has dramatically reduced the cost per sequenced base by several orders of magnitude and increased accessibility for metagenomic data production (Fig. 1). Improved affordability has catalysed the launch of numerous global-scale metagenomics initiatives[4–7], which have been pivotal for discovering novel microorganisms and enhancing our understanding of how microorganisms interact with their environments.

Despite the advancements made for metagenomic studies, the absence of standardized protocols for sample processing currently limits reproducibility[8] as the quality and quantity of extracted DNA are greatly influenced by the methods used for sample collection and DNA extraction[8–10]. Contamination poses an additional challenge, particularly in low-biomass samples[11] or those affected by environmental factors such as extreme climates, in environmental metagenomics studies[12]. The choice of sequencing technology further constrains sample preparation[13,14]. Additionally, the lack of consistent standards across reference databases leads to incomplete and inconsistent records[15,16]. Enhancing experimental protocols, computational methods and analytical approaches remain active areas of research, and efforts are underway to standardize the collection and processing of diverse metagenomic samples to mitigate contamination, address issues with low-biomass samples and ensure more robust metagenomic analyses.

In this Primer, we offer an overview of essential metagenomics concepts and describe current practices for the generation and analysis of metagenomics data, the applications of metagenomics, initiatives to improve metagenomic research and the future scope of the metagenomics field. We also review the current limitations of the experimental and computational methods used in metagenomic studies and discuss possible strategies to mitigate these issues. Finally, we emphasize the need for the synergistic design of the experimental and computational components of metagenomics experiments, as well as collaborations between basic science and applied practice, to facilitate novel applications of metagenomics in areas such as environmental monitoring, agriculture, biotechnology and medicine.

Experimentation

The choice of an appropriate sample collection method, preservation technique and sequencing workflow is necessary to form a reliable foundation for downstream metagenomic preprocessing and analysis. Given that each step can introduce bias[17], the entire process from sampling through laboratory analysis should be carefully designed and executed[18–22]. Different standardized protocols are required depending on whether microbiome samples are sourced from host-associated or environmental sources. Collected samples must be handled carefully before being transported and stored diligently to maintain the integrity of metagenomic information prior to DNA extraction. After DNA extraction, library preparation steps including DNA fragmentation, end repair, adapter ligation and sequence indexing are required to render the sampled DNA compatible with the sequencing platform being used. We discuss the experimental protocol to prepare a metagenomics sample for bioinformatics analyses below; for an overview of the experimental workflow, see Fig. 2.

Study design

A rigorously designed study maximizes the potential for generating high-quality, reproducible data and ensures that the resulting conclusions are scientifically robust and defensible. The study design must be carefully aligned with specific research objectives, as it directly impacts the validity and interpretability of results. Researchers should select the most appropriate design based on the study's goal, which can range across human, animal and environmental studies. Observational studies monitor a subject without intervention to reveal natural microbiome variations. Cross-sectional studies provide a snapshot of the microbiome at a single time point, which can be useful to find associations between the microbiome and specific outcomes (such as health outcomes). Case–control studies compare microbiomes between individuals with a condition and those without, identifying condition-linked differences. Longitudinal studies track microbiome changes over time, offering insights into temporal dynamics. Finally, randomized controlled trials — the gold standard for causal inference — rigorously assess the effects of specific interventions by minimizing biases. After selecting the appropriate study design, it is critical to estimate the necessary sample size that will give sufficient statistical power to detect any expected effects. This process can be achieved through a power analysis, using effect size estimates from pilot studies or existing literature[23,24].

Microbiome sample collection

Several factors should be considered for sample collection: whether the sample is representative of the study's objective; the potential for contamination from extraneous sources or cross-contamination; the practicality of sample collection, such as sampling costs, convenience and efficiency; and the nature of positive and negative controls. We consider these aspects below for various different types of microbial community samples.

Human microbiome samples. The majority of human microbiome studies have analysed the gut microbiome using faecal specimens owing to their large bacterial biomass and ease of collection[9,10], although many other tissue types and specimens have been investigated, including human milk[25], tumours[26], the respiratory tract[27], the vaginal environment[28,29], the urinary tract[30], skin[31] and saliva[32]. Priorities when collecting these samples include maximizing patient comfort and compliance while also ensuring robust preservation of

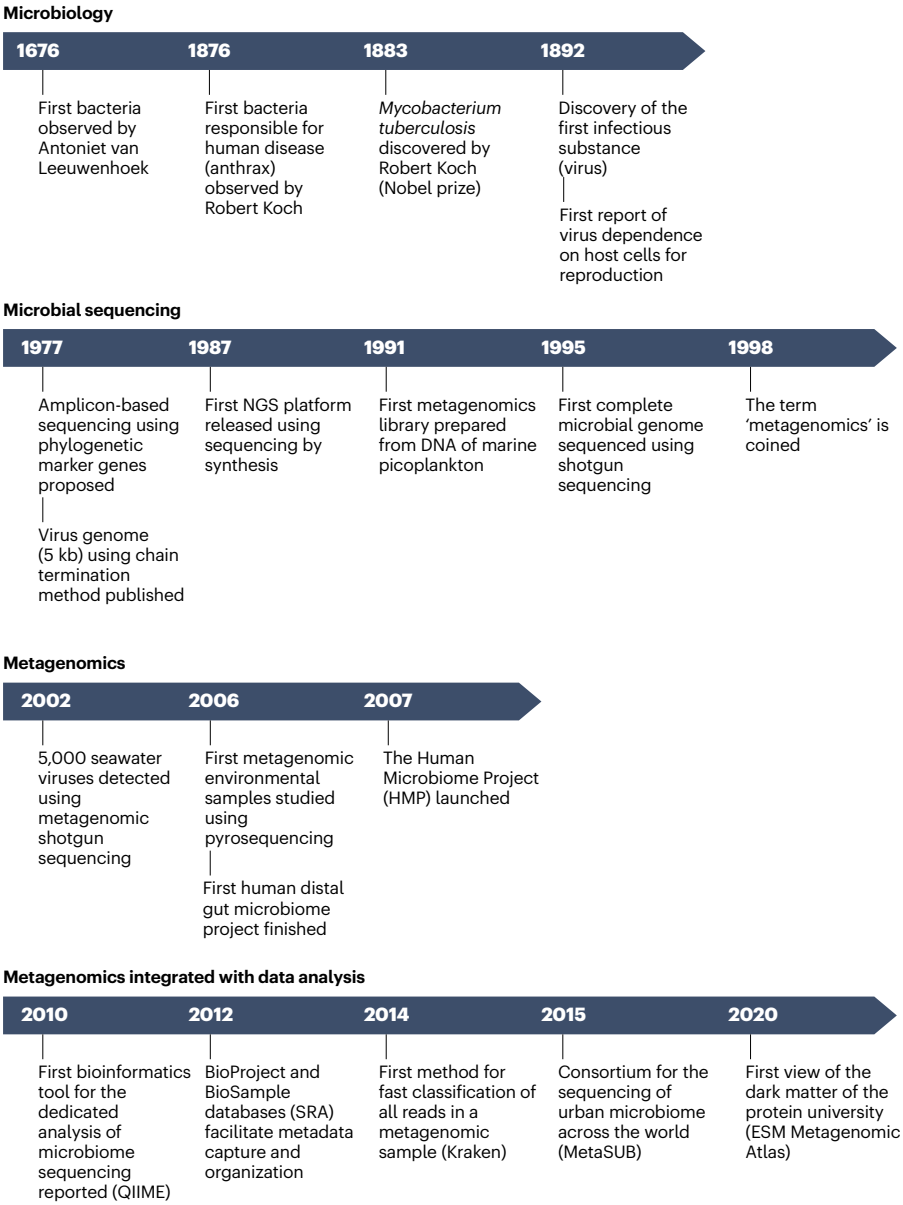

Fig. 1 | Timeline of microbial discovery and the development of metagenomic analysis. From the findings of early microbiology and advancements in sequencing technology, the field of metagenomics has grown to encompass the development of metagenomic databases, tools and organizations dedicated to the field of microbial and metagenomic discovery. NGS, next-generation sequencing.

bacterial communities. Challenges of both inpatient and outpatient sample collection include the need for freezing methods for sample preservation, contamination, low-biomass samples and availability of hospital resources[33]. In the context of faecal specimens, studies evaluating aspects of collection protocols such as the stabilization of samples and freezing methods have reported varied results on how DNA composition is impacted by different preservation methods; however, freezing freshly collected stools remains a gold standard to yield a reliable taxonomic resolution[16,33]. Commercial products are available for at-home stool collection; self-collection of the first bowel movement of the day is usually recommended[16] along with immediately freezing the sample or storing it in a preservation solution[34].

Non-human animal samples. In animal studies, sample collection varies depending on the study's objective, the tissue or specimen of interest and the host species. Faecal sample collection, rectal swabbing and post-mortem sampling of the intestinal contents of the host animal are options for gut microbiome studies. There is debate in the field about the interchangeability of sampling strategies. For example, in birds and reptiles there is an ongoing debate regarding the suitability of swabbing the cloaca to sample the gut microbiome[35,36], given the convergence of the urinary, reproductive and gastrointestinal tracts at that site. In mice and rats, faecal pellets and caecal contents collected at sacrifice are common samples for analysis. The non-invasive collection of faecal pellets facilitates repeatable longitudinal studies preventing bias in statistical interpretations[37], whereas caecal contents yield a more comprehensive profile of the gut microbial community but require euthanization[37]. Therefore, considering the experimental question and the type of study required is important when selecting the appropriate sampling method.

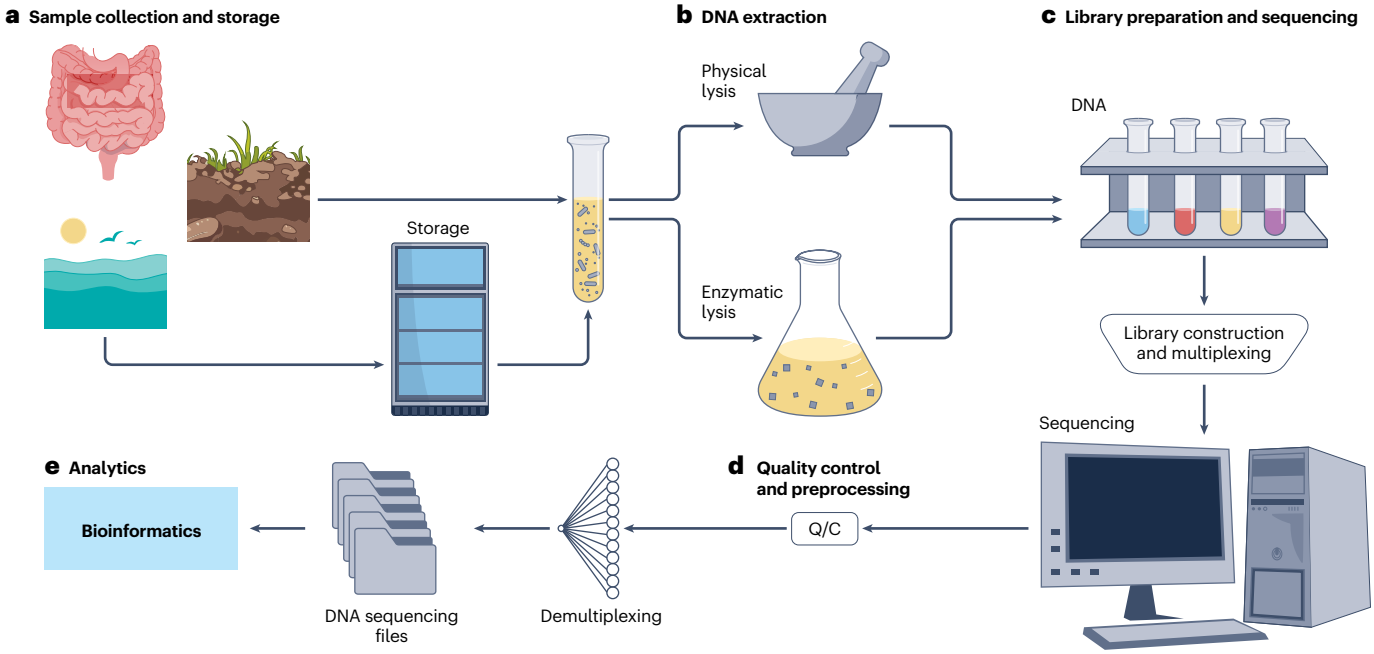

Fig. 2 | Experimental protocol for metagenomics experiments. **a**, Metagenomic samples, such as environmental samples from soil or water, or samples from the microbiome of organisms, are collected and either stored or processed immediately. **b**, DNA from the sample is extracted using physical or enzymatic lysis. **c**, A DNA library is constructed. Multiple samples can be sequenced together (multiplexed) by labelling samples with DNA barcodes. Bulk DNA is sequenced using whole shotgun sequencing. **d**, Sequencing reads undergo quality control checks and preprocessing. **e**, Demultiplexing separates sequences by DNA barcode, and sequences are then processed using bioinformatics analyses.

Environmental samples. Environmental samples for metagenomics include air samples, swabs of surfaces, and water and soil samples. Air sampling typically involves passing air through filters of various sizes to enrich for airborne microorganisms[38–40]. This method is particularly challenging because of the low bacterial biomass of the resultant samples[41] and because weather conditions such as wind, temperature or humidity can influence the microbial community yielded by this method[42]. Surface samples, such as samples of frequently touched surfaces in urban regions, can be collected using flocked swabs[6]. The composition of microbial communities obtained by this method can be influenced by variables such as surface type, cleaning frequency and human activity[43]. By contrast, water samples should be collected at a pre-specified depth and filtered upon collection[44,45]. Soil samples should be collected using sterilized standard tools and stored in a sterile bag. Soil sampling should consider the spatial and temporal variability of the sampled soil environment. For example, the density of microbial populations in the topsoil and subsoil differ, resulting in profile variations between these sampled communities. It is therefore important to note that microbial profiles can vary greatly across sample locations and seasons. To control for these variations, standardizing sampling protocols to sample across multiple time points and locations can capture natural microbial fluctuations; further, adjusting for environmental metadata (such as temperature, humidity or nutrient concentration) can further ensure that observed differences reflect true biological or environmental variations. Our recommendation is for researchers to choose a sampling method that minimizes the effects of contamination by maintaining a consistent sampling strategy throughout the study and to collect negative control samples, such as swabs of sampling equipment, air or areas adjacent to the sampled area, depending on the method used.

Sample handling, transport and storage

As microbiome communities are living and metabolically active meta-organisms, samples should be frozen or inactivated immediately after collection and then maintained under stable conditions until DNA extraction. Temperature fluctuations[34,46], oxygen exposure and multiple freeze–thaw cycles[46] should be avoided as these can compromise DNA and RNA integrity, affecting the metagenomic profiles of microbial communities[47]. Multiple freeze–thaw cycles can also bias sequencing results by selecting for harder-to-lyse bacteria.

The gold-standard method for sample handling after sterile collection is snap-freezing in liquid nitrogen, followed by storage at −80 °C. However, this is often not possible in field studies and many alternative methods for sample stabilization and storage have been evaluated[48–51]. These include preservation buffers such as RNAlater, ethylenediaminetetraacetic acid (EDTA) and ethanol-based or guanidine isothiocyanate-based[52] stabilizing agents, which are available in various commercial kits. When using stabilizing reagents, it is imperative to maintain a proper stabilizer to sample ratio and to thoroughly mix the stabilizer and sample. The choice of stabilizing solution, storage time and storage temperature can affect sample degradation, even for samples that were transferred immediately to −80 °C. For example, the efficiency of stabilizers can fluctuate between metagenomic samples, influencing the identified taxonomic composition of a microbial community as shown in faecal specimens[53]. Additionally, prolonged storage[54] or fluctuations in temperature[46] can potentially compromise the integrity of the microbial community and bias downstream analyses, as has been seen in pig faecal[55] and sewage samples[56]. Thus, the optimization of methods used to ensure stability in metagenomic samples is important to ensure reliable interpretation of microbial communities.

To ensure reproducibility, details of preservation methods and storage temperatures should be included in study metadata when possible. Additionally, this information should be uploaded to data repositories, such as protocols.io[57] and STAR methods[58], to allow for broad use by the research community[17]. Freeze–thaw cycles should be avoided by aliquoting samples prior to long-term cryopreservation, and appropriate negative and positive controls should always be included for data accuracy and potential bias quantification. Although bias due to preservation methods is unavoidable, careful consideration of sample storage and downstream bioinformatics pipelines can minimize its overall impact on the research objective.

DNA extraction

DNA extraction (Fig. 2b) involves the lysis of microbial cells and the isolation and purification of the now-accessible DNA. Among all sample preparation steps, the DNA extraction methodology might have the greatest impact on study variability due to unique biases in different microbial lysis approaches[18,59–61]. The type of DNA extraction method used is dependent on the type of sequencing platform that will be used (short-read or long-read). A bead-beating step should be used for the extraction of samples for short-read sequencing to ensure the effective lysis of Gram-positive bacteria and fungal cells and reduce the hands-on extraction time[21,50,60–63]. Conversely, enzymatic lysis techniques should be used for samples for long-read sequencing as mechanical lysis methods such as bead-beating result in DNA shearing[64,65], although there are still unresolved issues with respect to enzymatically lysing fungal cells[66]. Recent advancements have enabled the production of long reads from DNA extractions, which has led to the development of novel protocols[64] and benchmarking studies[54,66,67] that take these new capabilities into account. DNA extraction kits or methods such as phenol–chloroform or CTAB extractions use different techniques to increase the DNA yield and remove enzyme inhibitors from their samples. Benchmarking studies have been performed within and across different environmental samples in order to identify the best DNA extraction approaches for short-read metagenomics[51,59,61–65]; although some studies recommend particular kits or approaches for certain types of environmental samples, no single kit or approach has been identified as best.

Incorporating negative controls is crucial for detecting contamination across sampling and DNA extraction stages, which might otherwise lead to erroneous background detection of microorganisms and antibiotic resistance genes[66]. Using a field blank (filling a sampling tube with molecular-grade water) effectively captures potential contamination sources within the workflow. Additionally, positive controls (such as a mock community of bacteria processed as a separate sample) or internal standards (whole cells or exogenous DNA or RNA added to the sample matrix) should be included to identify potential biases in sample concentration, DNA extraction and bioinformatics analyses.

Library preparation and sequencing

After DNA extraction, the library preparation process (Fig. 2c) converts raw DNA extracts into labelled libraries ready for sequencing on the platform of choice[67]. Short-read library construction generally includes four main steps: DNA fragmentation into smaller, random, overlapping fragments using either sonication or enzymatic methods[68]; fragment end repair; ligation of pair-end or single-end adapters; and indexing[69]. Indexing allows multiple samples to be combined in a single sequencing run, reducing the monetary cost per sample and increasing the throughput per run. Indexed libraries are normalized and pooled to equimolar concentrations prior to sequencing[70].

Sequencing platforms. Illumina NGS platforms are often used for sequencing in metagenomic studies for their high-throughput and high-accuracy generation of short reads (ranging from 50 to 300 bp). These platforms are highly parallelized, allowing for the simultaneous sequencing of many DNA fragments, and have been successfully applied to shotgun metagenomics research to date[70]. These platforms are ideal for tasks such as variant detection, gene expression analysis and metagenomic profiling, where throughput and cost are paramount considerations. The established infrastructure surrounding short-read technologies, including robust analysis pipelines, further strengthens their appeal for high-throughput applications.

Long-read sequencing technologies such as Nanopore (Oxford Nanopore Technologies (ONT)) and PacBio (Pacific Biosciences) can produce reads several hundred kilobases in length and have been successfully applied for metagenomics research. These approaches allow real-time data acquisition as they can generate ultra-long reads that can be used to generate full-length mRNA or viral sequences. This allows direct sequencing from native DNA, reducing error rates and allowing solving complex region sequences. Long-read sequencing excels in providing comprehensive genome assemblies, accurate mapping of complex regions and identification of large structural variants. This technology can sequence native molecules without amplification, thereby avoiding PCR-induced biases and preserving epigenetic modifications, making it valuable for specific applications such as de novo assembly and isoform identification. For example, when investigating the genetic context of antibiotic resistance genes and their association with mobile genetic elements or host organisms, long-read sequencing offers superior accuracy owing to the reads spanning larger genomic regions[71,72], thereby reducing the bioinformatics biases commonly introduced by short reads[73]. Despite these advantages, Nanopore sequencing platforms exhibit a higher average base error rate (4–10%) compared with Illumina technologies (approximately 0.1%)[74] and Illumina platforms generally outperform Nanopore sequencing in terms of read yield. Advancements in long-read sequencing technologies such as the Nanopore PromethION sequencing platform[75,76], the R10.4.1 flow cell and HiFi sequencing using the PacBio Sequel II system have been used to address limitations regarding the sequencing depth, error rate and overall sample coverage in metagenomic research[77].

When selecting a sequencing platform, researchers must consider the goals of their study, the level of genomic complexity of their samples and the balance between cost-efficiency and data accuracy. The cost disparities between short-read and long-read sequencing can be substantial, with the cost of long-read platforms typically increasing several times per gigabyte of data generated[71]. This difference arises from the longer duration of runs and the greater complexity of library preparation and sequencing chemistry for long-read sequencing[71,72]. Consequently, although long-read sequencing provides deeper insights into genomic architecture, its higher per-unit cost can be prohibitive for large-scale projects. To avoid these limitations, hybrid sequencing approaches have gained traction by integrating the accuracy and cost-efficiency of short reads with the long-range continuity of long reads, thus offering a practical compromise and enabling high-quality genome reconstruction at a reduced cost[73,77].

Library preparation and sequencing. Library preparation for Illumina sequencing can be classified into PCR-based and PCR-free methods. PCR-free methods typically require a minimum DNA input of approximately 25 ng, whereas PCR-based library preparation is flexible in terms of DNA input, although it can introduce biases due to the

PCR amplification process. After selection and the preparation of the library, adapter-ligated DNA molecules are immobilized onto the surface of a flow cell, which is coated with oligonucleotides complementary to the adapters. These DNA molecules are then amplified through a process known as bridge amplification[77–79] and sequenced using an optical detection method. Illumina's NGS platforms offer single-read sequencing, which sequences DNA from one end of a fragment, and paired-end sequencing, which captures both ends of a DNA fragment[80].

The effectiveness of long-read sequencing heavily depends on the library preparation step. Over-shearing during preparation can compromise read length and quality. Specialized kits are available for library preparation. In ONT's DNA by ligation, Rapid and 16S library prep kits, DNA is sheared to fragments longer than 8 kb, end-repaired and ligated with protein–conjugate adapters, followed by a conditioning step before sequencing[81]. PacBio's SMRTbell library prep kit involves ligating universal hairpin adapters to sheared or amplified DNA fragments, preparing them for high-fidelity, long-read sequencing[82].

One of the main advantages of long-read sequencing options such as nanopore sequencing and PacBio's single-molecule real-time sequencing is their ability to produce reads that span complex genomic regions. In nanopore sequencing, bases are called when DNA passes through a nanoscale protein pore, generating fluctuations in electric current from which canonical and modified nucleotides can be identified in a context-dependent manner[83]. By comparison, single-molecule real-time sequencing has two modes: circular consensus sequencing for highly accurate long reads (HiFi reads); and continuous long-read sequencing for long reads where half of the reads exceed 50 kb in length[84,85]. These technologies offer considerably longer read lengths than Illumina's NGS platforms, although they come with trade-offs, such as the context dependency of base calls, which leads to a higher error rate for nanopore sequencing, and the need for repeated base calling to build a consensus, which leads to higher costs for single-molecule real-time sequencing.

A variant of paired-end reads can be generated using high-throughput chromosome conformation capture (Hi-C), a technique initially developed to study the three-dimensional organization of chromosomes in which paired reads serve to link genomic regions that are in proximity within the cell. This arrangement results in some Hi-C pairs spanning extensive genomic regions as distant DNA segments come into close proximity within the cellular milieu. The combination of shotgun sequencing and Hi-C has led to the development of metagenomic Hi-C (metaHi-C)[86,87], which involves shotgun extraction of genomic fragments from a microbial sample, along with a Hi-C experiment generating DNA–DNA proximity ligations between loci within the same physical cell, thus enabling the linking of contigs assembled from the shotgun sequencing.

Sample multiplexing is a common practice in metagenomic data preprocessing for conducting high-throughput and cost-effective analyses of several microbial communities. A metadata file containing details of individual indexes and their corresponding samples is essential for accurate demultiplexing. In addition, indexes should be designed based on error-correcting bioinformatics methods such as CD-HIT[88], DNACLUST[89] and Shepherd[90] to mitigate misassignment of reads due to sequencing errors[79].

Results

The initial stages of metagenomics data analysis include preprocessing and quality control steps, including assessing read coverage, depth analysis in relation to a reference genome, assessment of contig completeness and additional determination of contamination levels in assembled contigs. The bioinformatic analysis of metagenomic data then moves onto more computationally demanding tasks such as metagenomics assembly and functional analysis of microbial community, which require advanced computational infrastructure such as high-performance computing clusters for efficient data processing. Here, we explore an array of strategies and bioinformatic tools commonly used for efficient data processing and metagenomic analysis, and discuss challenges and mitigation strategies to obtain reliable results effectively. Commonly used analytical tools for metagenomic analyses are presented in Supplementary Table 3.

Quality control of metagenomic data

Before quality control, a demultiplexing step is essential for distinguishing individual samples based on the unique indexes incorporated into the reads during library construction. Tools for demultiplexing have been developed, including Flexbar[91,92] and Ultraplex[93], among others[94–99].

The stochastic process of whole-genome shotgun sequencing on metagenomic samples can introduce biases when performing metagenomic assemblies, making it challenging to accurately reconstruct genomes within a sample. Therefore, it is important to assess the quality of the raw sequencing data after demultiplexing. Read quality is measured using the Phred quality score, provided by the sequencing platform used, which is based on the probability that a base was sequenced incorrectly. Low-quality reads — those with a Phred score <30 — often contain technical artefacts such as sequencing errors, PCR artefacts and adapter sequences, and should be eliminated using quality control and data clean-up tools such as FastQC, PRINSEQ[100], Trimmomatic[101], BBTOOLS and others[102]. These tools leverage both the quality information provided by sequencing instruments and databases containing collections of adapters and primer sequences to facilitate the removal of sequencing adapter sequences and investigate read sequences for anomalies, such as the over-representation of *k*-mers (a sign of contamination or genomic repeats that can lead to complex metagenomic assemblies). They also assess the read length, quality scores, GC content and number or percentage of ambiguous bases. Post-trimmed reads that fall below a predefined threshold length are discarded.

Alternative quality control methods are used for long sequencing reads[103–106]. For example, the NanoFilt tool filters long sequencing reads based on common parameters such as average read quality, length and GC content[104], and LongQC[103] assesses the quality of long reads by evaluating the proportion of atypical reads — known as 'nonsense reads' — potentially generated from low-quality pores. Reads derived from the host genome or from other unwanted sources of contamination must be removed; automated decontamination tools such as DeconSeq[105] and KneadData streamline this process.

Metagenomic assembly

Analyses can be performed directly on individual reads derived from metagenomic samples; however, the assembly of reads into individual contigs and metagenome-assembled genomes (MAGs) is often preferable as this provides higher resolution (more detailed and accurate differentiation among species and strains) of the genomic content of microbial communities and enables improved functional annotation[102] (Fig. 3, left side).

Contig assembly. Metagenomic assembly tools largely employ extensions of assembly algorithms developed for isolated genomes that

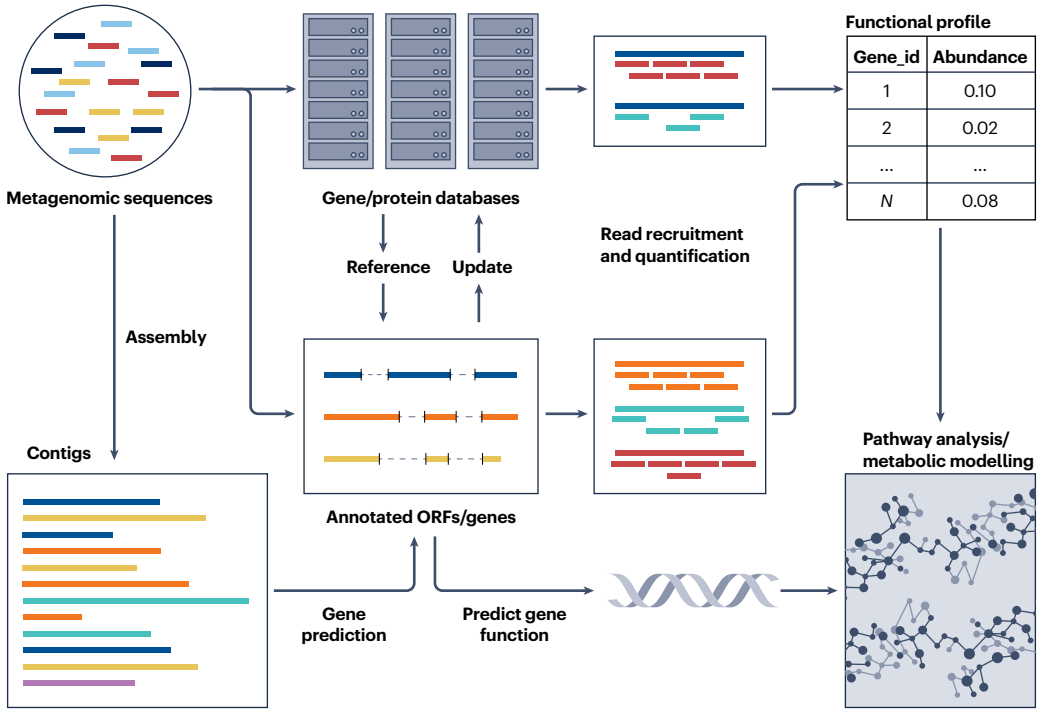

Fig. 3 | Metagenomic functional analysis. Metagenomic sequences are a common starting point for functional profiling and annotation (left side). Known genes can be directly aligned to a reference for alignment-based quantification (top row). Metagenomic sequencing reads can also be employed by ab initio gene finders to train models that identify open reading frames (ORFs) from input data followed by read quantification (middle row). Both alignment-based and ab initio methods can be used for functional profiling, which can be further analysed to explore biological insights (right side). Metagenome sequences can be assembled into contigs to facilitate gene discoveries by identifying novel ORFs and predicting gene function based on existing knowledge (bottom row).

have been adapted to the uneven coverage and species/strain diversity commonly found in metagenomic data. The majority of short-read metagenomic assemblers rely on a de Bruijn graph — an approach that constructs a graph composed of *k*-mers extracted from the reads[107].

Genomic repeats create ambiguities that make it difficult to identify the path corresponding to the correct reconstruction of the genome. In single genomes, repeats can be identified and even resolved by analysing the depth of coverage within the assembly (the number of times each *k*-mer is seen in the reads). Short-read metagenomic assemblers have been extended and adapted to long-read sequencing technology. The metaFlye assembler[108] computes *k*-mer frequencies within a local neighbourhood in order to account for the varied coverage within metagenomic samples, in contrast to the long-read single-genome assembler Flye[109]. Similarly, hifiasm-meta[110] extends the hifiasm tool[107] by adjusting the criteria for removing short and chimeric reads to account for the potentially low coverage of some genomes in a sample. hifiasm-meta and hifiasm were specifically developed for the HiFi technology from Pacific Biosciences. Another approach developed for the assembly of metagenomic data from high-quality long reads is metaMDBG[111], which builds the assembly around minimizers identified in reads. Minimizers[112] are a strategy for summarizing genomic data in a small memory footprint, thus offering speed and memory advantages over de Bruijn graph approaches.

Hybrid assembly approaches combine the precision of short reads with the extended coverage of long reads and offer a powerful solution for metagenomic assembly. In a recent study assessing ten environmental metagenomic samples and in silico-generated data using seven assemblers (IDBA-UD[113], MEGAHIT[114], Canu[115], metaFlye[108], Opera-MS[116], metaSPAdes[117] and hybridSPAdes[118]), the hybrid approaches (Opera-MS and hybridSPAdes) consistently outperformed the others in terms of accuracy[119]. Benchmarking tools for hybrid assembly revealed that Unicycler exceeds MaSuRCA and SPAdes in producing contiguous genomes, particularly when combining Illumina and ONT data[120].

Despite advancements with hybrid assembly approaches, challenges remain in using error-prone long reads for reconstructing high-quality genomes from complex metagenomes (for example, those isolated from sludge or wastewater samples). A novel haplotype-resolved hierarchical clustering-based hybrid assembly (HCBHA) approach has been developed for this application[121]. This method phases short and long reads into distinct haplotypes before assembling each bacterial genome individually, enabling the reconstruction of near-complete genomes from highly complex ecosystems.

Metagenomic samples can contain multiple strains of an organism, each differing from each other in just a few genomic locations[122]. In this case, specialized assemblers similar to tools that haplotype the human genome[123] are required to assemble contigs from each strain. In contrast to the human genome, which has two haplotypes, the number of variants of a genome within a sample is not known a priori and is frequently larger than two. Initially, techniques for microbial community haplotyping have been developed in the context of viral quasispecies[124]. We discuss the general workflow of these strain resolution approaches in Supplementary Box 3.

We note that the multi-haplotype problem is an example of an 'NP-hard'[125,126] problem, where algorithms for solving it efficiently are

unlikely to exist and all practical approaches rely on heuristics that cannot guarantee a correct answer. Such heuristics might miss strains or create a mosaic composed of combinations of haplotypes that do not actually occur in the sample being analysed. Further challenges arise from sequencing errors and the fragmented and incomplete nature of metagenomic sequencing data. An alternative is to define haplotypes or strains only with respect to a set of conserved genes, an approach that does not require a high-quality assembly[127].

Contig binning. Contigs produced by a metagenomic assembler can originate from many different organisms and cannot be separated out a priori. These contigs can be organized into bins intended to represent individual organisms or taxa, relying on information that was not used during the assembly process. Contigs that have a similar coverage depth within a single sample, a similar coverage profile or a similar *k*-mer composition are usually assumed to be derived from a single genome.

Metagenomic binning tools typically annotate the contigs with a set of features (such as coverage and sequence composition) and then use clustering algorithms to construct the bins. For example, MetaBAT[128] uses a modified *k*-medoid approach to cluster contigs based on a distance metric that combines tetramer similarity with abundance information. CONCOCT[129] also incorporates both contig abundance and sequence composition information within the context of a Gaussian mixture model to identify clusters and bins. The more recent VAMB[130] relies on variational auto-encoders (a type of neural network) to cluster together contigs that are initially represented as points in a high-dimensional space. The embedding of contigs into this space is defined by their abundance across samples and their sequence composition. Some binning tools, such as MaxBin[131] and Binnacle[132], also incorporate mate-pair information in addition to coverage and sequence composition.

Bins that appear to represent nearly complete reconstructions of genomes from a metagenomic sample are known as MAGs to differentiate them from genomes reconstructed from cultured isolates. For details on the validation of MAGs, see Supplementary Box 4.

Taxonomic characterization in metagenomics

Taxonomic classification and profiling involves identifying and estimating the relative abundance of known microbial taxa in a metagenomic sample using their sequenced genomes as references. This approach requires sequence data for the taxa of interest and differs from the metagenomic assembly approach above, which aims to discover novel microbial genomes from DNA fragments in the sample. Both taxonomic classification and profiling compare sequencing reads or contigs[133] against a reference database, yet they differ in approach and purpose. Taxonomic classification uses taxonomic binning to assign individual reads or contigs to specific taxa, subsequently aggregating these assignments to estimate the relative abundances of taxa present in a sample. On the other hand, taxonomic profiling reports the relative abundances of taxa by comparing the overall sequence content of a sample against reference sequences. We discuss the methodologies for taxonomic classification and profiling of metagenomes and existing tools below.

Alignment-based strategies. Alignment-based methods for taxonomic classification rely on algorithms to align sequencing reads against reference genomes (Fig. 4Bb,Bc). Modern versions of these algorithms generally employ a taxonomic binning approach to determine the potential location of each read on a reference genome, assign a score evaluating the quality of the alignment and assign taxonomic labels to the aligned reads based on sequence similarity to the reference genome (Fig. 4Bd).

Alignment strategies using entire reference genomes represent the earliest algorithms used for taxonomic classification. An example is MEGAN, which aligns each read using BLAST and assigns them to their lowest common ancestor (LCA) — the ancestral node shared across a group of species from which the read may be derived[133]. The LCA is used because metagenomic read data often comprise short, fragmented sequences that might not align perfectly with known reference genomes, or might align to multiple reference genomes owing to high genomic similarities between closely related species or strains. By determining the LCA shared by multiple reference sequences that match a given query sequence, the LCA approach provides a robust method for taxonomic classification, effectively handling incomplete or divergent sequences, and can also be used in alignment-free or hybrid taxonomic profiling approaches (discussed below). The algorithm arranges less-conserved species closer to the 'leaves' and more-conserved species closer to the 'root' of a phylogenetic tree. Although MEGAN is optimized for short reads, the taxonomic binning strategy in MEGAN has been adapted to use long reads and assembled contigs as part of the MEGAN-LR[134] tool (Supplementary Box 5).

Curated subsets of unique genes with a single copy number can be used as marker genes to enhance the computational efficiency of alignment-based methods. The goal of marker gene-based approaches is to create a reference database from marker genes alone to reduce the size of the reference database and optimize resource use (Fig. 4Ba). This approach was able to reduce the RefSeq dataset from 9.8 TB to the MetaPhlAn dataset of 10.41 GB, representing a 98.95% reduction. Marker gene-based approaches can vary widely in the number of marker genes used, ranging from a few dozen to millions; for example, MetaPhyler uses 31 protein-coding marker genes chosen for their efficacy in phylogenetic analysis, spanning 581 genera, 214 families, 99 orders, 46 classes and 27 phyla[135]. Another marker-based tool, PhyloSift, uses 37 'elite' gene families alongside 4 supplementary sets encompassing 16S and 18S rRNA genes, mitochondrial gene families, eukaryote-specific gene families and viral gene families, totalling approximately 800 gene families, with a predominant representation of viral families[136].

The reduced reference database of marker genes gives a unique representation of specific taxa. The 16S rRNA gene, which is critical for bacterial and archaeal taxonomy, is ubiquitous across prokaryotic genomes and contains conserved regions that are essential for ribosomal function, as well as variable regions that allow for discrimination between different taxa. The V4 region of the 16S rRNA gene is frequently targeted for amplicon sequencing in microbial ecology studies for its high variability and species-level resolution, enabling researchers to identify microbial taxa and assess their relative abundances in environmental samples.

Some marker gene-based tools, including MetaPhlAn, align metagenomic reads to clade-specific marker genes to assess microbial relative abundances. The MetaPhlAn marker genes, which encompass both bacterial and archaeal phylogenies, were selected from more than two million potential candidates identified from available genomes, based on their high conservation within clades and minimal similarity to genes outside those clades. MetaPhlAn compares metagenomic reads against the clade-specific marker gene reference database using nucleotide alignments, thus enabling efficient estimation of clade abundances. MetaPhlAn4, an updated version

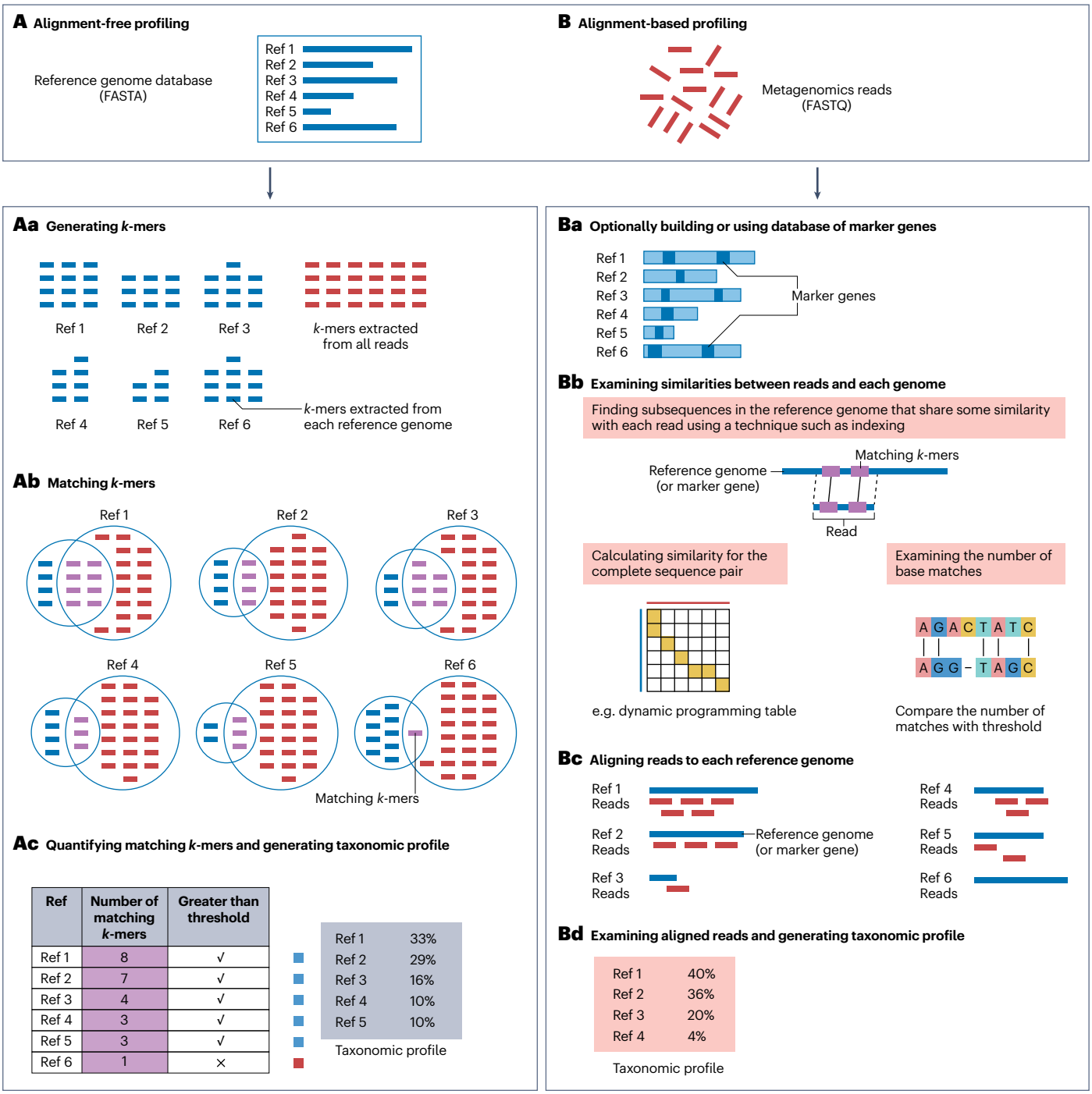

Fig. 4 | Metagenomic taxonomic characterization using alignment-based and alignment-free methods. In both alignment-based and alignment-free profiling approaches, a public reference database is chosen for the input of the metagenomic data. **A**, Alignment-free profiling tools in general experience reduced computational burden: the generation of *k*-mers is performed on both the chosen reference dataset and the metagenomic sample of interest (panel **Aa**); *k*-mers generated from the references and sample are compared to identify matching *k*-mers (panel **Ab**); and matching *k*-mers are quantified to generate a taxonomic profile (panel **Ac**). **B**, Alignment-based methods can be more sensitive at the cost of increased computational burden: the metagenomic sequencing reads are aligned to a prebuilt marker gene-based reference database (panel **Ba**); similarities between reads and each genome are calculated using either indexing, dynamic programming or *k*-mer matching (panel **Bb**); and reads are then aligned to the reference database (panel **Bc**) to be quantified for taxonomic profiling (panel **Bd**).

of its predecessor, introduces several enhancements, including the capability to construct custom marker gene databases from MAGs, enabling researchers to conduct more comprehensive and customized studies of microbial community diversity. Custom marker gene databases in MetaPhlAn4 are curated by combining marker genes from MAGs with those from reference genomes, thereby augmenting the accuracy and specificity of taxonomic profiling by incorporating genetic data from previously uncharacterized microbial taxa found in a metagenomic sample[137].

Marker gene-based strategies, such as those that use marker genes based on operational taxonomic units (mOTUs; universally conserved single-copy protein-coding genes present across diverse taxa) have been employed to improve the resolution of taxonomic profiling of microbial communities. The use of pairwise alignments for clustering marker genes has successfully identified 10 marker genes appropriate for creating more than 7,700 reference mOTUs for taxonomic classification and profiling[138]. Additionally, more than 20,000 reference mOTUs were recently added, identified through marker genes from approximately 150,000 MAGs[139]. It should be noted that marker gene-based approaches often exhibit increased specificity (true negative identification of species) over genome-based alignment strategies at the cost of decreased sensitivity (true positive identification of species), due to their dependence on conserved nucleic acid sequences to represent taxa[138].

Alignment-free, *k*-mer-based strategies. *k*-mer-Based methods for taxonomic classification generate *k*-mer profiles from metagenomic data (Fig. 4Aa), compare them against those of reference databases to identify matching *k*-mers (Fig. 4Ab) — allowing reads or contigs to be classified to a corresponding reference genome without the need for an exact genomic index — and quantify shared *k*-mers to generate a taxonomic profile (Fig. 4Ac). These alignment-free strategies are less computationally intensive than alignment-based methods. *k*-mer-Based methods provide more scalable solutions by incorporating *k*-mer usage in techniques such as exact matching[140], binning[130], hashing algorithms[141,142] and discriminative subsets[143] (for more information about *k*-mer-based machine learning approaches, see Supplementary Box 6).

k-mer matching has been used to map query *k*-mers against reference *k*-mers, where the LCA of the genomes containing each *k*-mer is used for taxonomic profiling[140]. This approach can present challenges; for example, the *k*-mer matching program Kraken uses a majority vote strategy for classifying reads, which might result in suboptimal performance at lower taxonomic rank levels due to high sequence similarities and sequencing errors. The Bracken tool was developed to overcome this sensitivity issue and probabilistically reassigns reads across the taxonomic tree, improving accuracy at finer taxonomic resolutions[144].

k-mer-Based sketching approaches provide a more computationally efficient alternative to *k*-mer matching. These methods condense large datasets into smaller 'signatures' while preserving the similarity relationships of *k*-mers, thereby minimizing the computational resources required for taxonomic profiling[141]. For details on specific *k*-mer based sketching approaches, see Supplementary Box 7.

Hybrid strategies. Hybrid approaches to taxonomic characterization offer a promising strategy for combining the strong precision and recall in taxonomic classification at the species and genus levels of alignment-based methods with the efficiency of alignment-free methods. For example, one hybrid approach[145] uses MinHash, a hashing technique that uses subsamples of *k*-mers to estimate genomic similarities, which is then coupled with alignment[146]. This combination allows Metalign to accurately profile metagenomic datasets by estimating the containment index — the likelihood that a given reference genome is present within a sample[146]. By reducing the size of the reference, this approach optimizes the alignment process for taxonomic profiling improving efficiency without compromising accuracy.

Short-read versus long-read taxonomic profilers. The use of short-read or long-read sequencing technologies has implications for the accuracy and resolution of taxonomic profiling. Long-read sequencing technologies can enhance taxonomic analyses by providing more comprehensive sequence information, improving the taxonomic resolution and accuracy of complex microbial communities[147]. A comprehensive benchmarking study assessing 11 taxonomic classification methods, including 5 designed specifically for long reads, showed that long-read classifiers generally outperformed short-read classifiers[127]. Short reads frequently produce false positives, particularly at lower abundances, and require additional filtering to achieve acceptable precision. By contrast, long-read methods such as BugSeq, MEGAN-LR and DIAMOND and the generalized method sourmash exhibit high precision and recall without requiring any filtering. Long-read methods also demonstrate superior performance for detecting species at low abundance levels; for example, these methods were able to detect species at 0.1% abundance with high precision in PacBio HiFi datasets, something that was more challenging for short-read methods.

Read quality can markedly impact the performance of long-read taxonomic profilers. Long-read datasets with a large proportion of shorter reads (<2 kb) resulted in lower precision and worse abundance estimates[127]. The study further demonstrated that long-read datasets provided superior taxonomic classification results compared with short-read datasets, particularly at finer taxonomic resolution, due to the greater contiguity of long-read assemblies. At both the species and genus levels, long-read datasets provided better detection metrics compared with short-read datasets[71,127].

Finally, the ability of ONT platforms to provide real-time sequencing reads, which can be used immediately for taxonomic classification without compromising precision or sensitivity due to the assembly and binning process, presents an advantage over traditional methods that require lengthy culturing and molecular diagnostics[148,149]. This is critical for metagenomic applications in clinical settings where accurate and timely pathogen identification is critical.

Functional analysis of metagenomics

Metagenomic functional analysis refers to the process of analysing genomic sequences from metagenomic datasets to decipher the functional capabilities and potential metabolic pathways within microbial communities in specific environments[150,151]. Functional analysis aims to uncover the capabilities and activities of these microorganisms, such as the biochemical pathways they use, the enzymes they produce and their roles in ecosystem processes. The first metagenomic functional analysis was carried out by cloning the functional genes isolated from environmental DNA into *Escherichia coli* and analysing enzymatic activity using non-computational methods[152]. This study was followed by others that demonstrated the presence of various microbial functions in environmental samples[3,153,154]. Large-scale studies of the human gut microbiome such as the Human Microbiome Project[155] have highlighted that microbial functionalities are more conserved across cohorts than taxonomic composition. The variability in taxonomic units observed

across cohorts can present a reproducibility challenge for comparative metagenomic studies, as differences in taxonomic composition can complicate the identification of consistent patterns across studies, suggesting a need to shift from taxonomic units to functional groups[156,157]. In this section, we discuss the functional analysis process.

Gene prediction approaches. Gene prediction approaches (Fig. 3, bottom row) identify potential genes from sequencing reads or assembled contigs. The key challenge of gene prediction is discriminating whether genomic regions that resemble genes are indeed authentic protein-coding genes; in complete genomes, such regions are open reading frames (ORFs), whereas in metagenomic applications the fragmented nature of the data requires the consideration of incomplete ORFs. Early alignment-based tools such as CRITICA[158] and Orpheus[159] infer gene presence by aligning query sequences to known protein databases (Fig. 3, top row); as a result, these approaches are only effective at detecting a subset of genes in a sample expressing previously characterized proteins. This limitation restricts their ability to uncover novel or unique genes due to their divergence from known sequences. More recent ab initio gene finders employ statistical models that are trained on known gene sequences to determine which ORFs are likely to be genes. These models analyse various features, such as *k*-mer frequencies, ORF length, GC content and sequence motifs to predict genes. Other methods such as hidden Markov models[160–165], support vector machines[166] and conditional random fields[167] have also been employed in microbial gene finders to improve the accuracy of gene prediction.

The parameters of ab initio gene prediction models can vary across bacterial species and complicate gene prediction in metagenomic samples as the organism to which each ORF belongs is typically unknown. Tools such as MetaGene and its successor, MetaGeneAnnotator[168], address this issue by focusing on species-specific features such as patterns of ribosomal binding sites and di-codon usage in training their probabilistic models. This enhances the precision of gene prediction across metagenomic data[168]. Prodigal, another widely used tool for microbial gene prediction, leverages the start codon and ribosomal binding site motifs, GC content and hexamer usage patterns to build a model that can identify and annotate genes in microbial genomes[169]. Recently, deep learning tools such as CNN-MGP[170] and Meta-MFDL[171] have been used for gene finding and demonstrated promising results.

Despite the numerous gene prediction tools available, no consensus exists on a definitive gold standard. A benchmark study demonstrated that ab initio methods generally surpass evidence-based tools, with performances varying across different organisms[172]. To further optimize gene prediction, the annotation tool MetaErg incorporates additional features such as signal peptides and transmembrane helices, which are implicitly contained in trained models, to enhance the filtering of predicted ORFs[173].

Annotating functions in a metagenome. Functional annotation in metagenomics (Fig. 3, left side) involves computationally assigning biological or functional information to metagenomic sequences, facilitating an understanding of the roles and activities of microbial communities within their environments. It essentially involves identifying and elucidating the functional potential of each gene, thereby expanding the knowledge base of gene functions and, potentially, discovering new genes or functions. Functional annotation complements functional profiling (covered in the next section) as it focuses not only on identifying known functions but also on predicting roles for previously unknown genes, whereas functional profiling quantifies the abundance of known genes in a sample (Fig. 3, middle row)[173–179].

Comparing unknown genes against reference databases is a fundamental step in functional analysis. Orthology-based approaches such as eggNOG-mapper[180] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) KOALAtool[181] leverage databases such as Clusters of Orthologous Groups (COGs)[182] and KEGG[183], respectively, to classify genes into evolutionarily related groups to facilitate gene function prediction using shared ancestry. Additionally, operational functional units, which group genes or gene products based on their functional similarities, are emerging as a new benchmark for understanding the functional potential of microbiomes[184]. Other annotation tools that use structural similarity and incorporate clustering or transfer learning approaches include MorF[185] and FunFams[186]. The tools ProtBERT[187] and ProSE[188] rely on sequence embeddings. Additional efforts have been made in the construction of protein families through databases such as Pfam[189] and AGNOSTOS-DB[190], which are instrumental in identifying proteins with related functions and elucidating unknown boundaries of functions.

Despite advancements in functional annotation tools, a significant fraction of microbial functions cannot be properly annotated[176]. To increase the proportion of sequences with identified functions, it is important to adopt gene context-based methods beyond traditional orthology-based approaches such as FunGeCO and integrate the genomic environment surrounding a gene, thereby enhancing functional annotation[184]. Alternatively, the subsystems approach, exemplified by MG-RAST[191], organizes gene families into subsystems within functional networks, thus enhancing our understanding of microbial metabolism. Moreover, deep learning-based methods, such as DeepFRI[192], represent a competitive alternative, effectively combining sequences with structural features to substantially increase the coverage of functional information.

Functional annotation of metagenome sequences can serve as a valuable tool to predict and discover novel genes involved in metabolic pathways and enzymatic functions, antibiotic resistance genes and CRISPR–Cas systems, valuable in different biotechnological applications[193]. Additionally, functional annotation enables the exploration of metagenomic functional signatures across various environments, such as nitrogen cycles[194], disease-specific microbiomes[195] and virus–host microbiome interactions[196].

Functionally profiling metagenomic data. Functional profiling elucidates the functions encoded by the annotated genes to understand the functional dynamics and ecological roles of the diverse microbiota (Supplementary Table 1). This approach involves aligning query sequences against databases of sequences with known functions or employing non-alignment approaches, such as machine learning models trained on sequences with known functions. Some approaches operate directly on reads, whereas others are applied to the sequence of genes identified in the samples. The latter approach offers advantages as the analysis is performed on longer DNA segments, which reduces redundancy and improves analytic accuracy.

Alignment-based functional profiling methods use sequence comparison algorithms to compare the predicted amino acid sequences of genes with sequences in a reference database. For example, MG-RAST[197] uses BLAST to search against the M5nr database[198], a curated collection of non-redundant protein sequences from the National Center for Biotechnology Information (NCBI), UniProt and KEGG. Additionally, BlastKOALA and GhostKOALA, which have been developed

specifically for the functional annotation of protein sequences in microbial genomes, perform searches against KEGG Orthology (KO) terms[181]. DIAMOND[199], which is specifically designed for short reads, uses a seed search approach and double indexing to improve speed and accuracy when querying against databases such as the NCBI and KEGG. Sequence alignments can enhance the capabilities of alignment free-based profilers; for example, eggNOG-mapper[180], which uses precomputed sequence clusters to facilitate comparisons, uses DIAMOND as one of its efficient sequence mapping options.

Tools using hidden Markov models, such as InterProScan[200], have been developed to facilitate comparisons of genes against orthologies using precomputed sequence clusters, which is less computationally intensive than searching through all sequences in a database as in alignment-based strategies. For example, KOfamKOALA[201] reduces the computational time of alignment-based programs by using HMMER/hmmbuild to produce profile hidden Markov models and implementing an adaptive score threshold to accurately delineate metabolic and regulatory functions.

Protein domains or motifs can be annotated to further evaluate the functional properties of microbial communities. Functional annotation can use protein domains to identify conserved regions and infer putative biological functions at a proteomic level, where proteins sharing similar domains are classified into protein families, providing evolutionary insights into microbial communities. Proteomic information can be found in databases that store validated protein domains and families, such as Pfam[189], UnitProt[202] and SWISS-PROT[203]; tools such as HMMER[204] and InterProScan[200] can be used to search for domains to assign putative functions to protein sequences.

Pathway-based methods enhance functional analysis by annotating the pathways within metagenomic data through the identification and interpretation of gene functions, as well as protein and molecule actions. Such information can be particularly valuable because pathways might span multiple organisms, be partially represented or even serve as the foundation for constructing mechanistic models. These methods, which rely extensively on well-curated reference databases, not only improve our understanding of metabolic functions but also underscore the complex interplay within microbial ecosystems. For example, HUMAnN (HMP Unified Metabolic Analysis Network) annotates and reconstructs human metabolic pathways[205] by mapping sequencing reads to reference genes or protein families that constitute known functional pathways, enabling the exploration of ecological and biogeochemical functions between the host and the environment. Newly developed tools such as gutSMASH[206] profile known and predicted novel metabolic pathways by using metabolic gene clusters to conduct functional comparisons in different cohorts. Pathway-based databases include MetaCyc[207], KEGG, KBase[208] and DrugBank[209].

Applications

Potential applications of metagenomics are broad, spanning from clinical applications such as pathogen surveillance and monitoring antibiotic resistance to applications in ecology such as studying the crucial ecological roles of microorganisms, analysing biodiversity, and assessing ecosystem functions and the impact of environmental changes. We discuss these applications below.

Microbiome–disease associations

Bioinformatic analysis of metagenomic sequencing data has become an important tool in understanding shifts in human microbiome composition associated with various health and disease states. This approach enables the precise identification of specific microbial taxa and functions linked to health conditions and the exploration of the complex dynamics of microbial interactions during disease progression.

Researchers often employ either a cross-sectional or a longitudinal study design to understand the association between diseases and the microbiome. In a cross-sectional study design, a cohort with a specific disease is compared with a healthy control group. Characteristics of the microbiota, such as taxonomic composition and functional profiles, are analysed and compared between groups to identify differences that might correlate with health status. Typically, the relative abundance of different taxonomic groups is evaluated to determine whether specific taxa correlate with disease status and statistical tests, and standard *t* tests and complex linear models are used to assess associations or differences between groups[210]. These techniques have been used successfully to find associations between the microbiome and both diseases of the digestive system and systemic disorders[210]. For example, a decrease in butyrate-producing bacteria and an increase in functional groups responsible for sulfate reduction and oxidative stress resistance is seen in patients with type 2 diabetes[211]. Similarly, in major depressive disorder, numerous pathways involved in amino acid metabolism were shown to be disrupted in the faecal microbiome[212]. In rheumatoid arthritis, both the oral microbiome and the gut microbiome are altered, with functional differences in the redox environment and the transport and metabolism of iron, sulfur, zinc and arginine[213].

Longitudinal study designs involve sequencing the microbiota at various time points during the progression of the disease to track changes in microbial composition over time and obtain insightful information into microbiome dynamics changes alongside disease progression. For example, longitudinal studies on the oral, lung and gut microbiota of patients with acute respiratory failure and lower respiratory tract infections performed across different hospitals have revealed distinct patterns of taxonomic and functional profiles; these studies then evaluated changes in these patterns in response to disease progression and treatment[214,215].

Measuring microbial diversity can aid in revealing compositional differences in microbiota across different environments and samples. Statistical measures such as Shannon's diversity index[216] or bioinformatics tools such as phyloseq[217] have been developed to study microbial diversity in metagenomic samples. For example, a study by Yin et al. used the Shannon index to evaluate the impact of gut microbiome diversity on mortality risk in patients with septic shock[216]. Using these statistical methods, researchers have identified correlations between low oral microbiome diversity and periodontitis[218], as well as between low gut microbiome diversity and various ailments, ranging from prediabetes[219] to Crohn's disease[220]. The practical application of microbial diversity is often limited to conditions where the microbiome's impact is well understood, and the use of alpha diversity indices[221,222] to study the impact of the microbiome on health conditions such Parkinson disease, multiple sclerosis and certain forms of depression has proved inconclusive results.

Despite significant taxonomic variability, the human gut microbiome often maintains a conserved functional profile[223]. Analysing metagenomic data at the gene and pathway levels can help researchers elucidate the mechanisms through which the microbiome interacts with host factors in disease processes. This analysis holds promise for clinical applications in both diagnosis and treatment. For example, the recent development of q2-predict-dysbiosis, a metagenomics-based

gut health evaluation tool described in a recent preprint article[224], leverages microbial functions and interactions and has shown significant improvements over most traditional taxonomy-based indices for the diagnosis of specific health conditions such as inflammatory bowel disease.

Functional analysis relies on the completeness and accuracy of functional units, such as genes, gene clusters or protein families, which require proper categorization and distinction within samples by utilizing information on genes and pathways, as well as assessing their abundance levels. Bioinformatic tools such as Anvi'o[225] or Picard can be used for this purpose for clustering contigs based on sequence similarity and annotating genes and metabolic pathways using reference-based alignment.

Clinical diagnosis using metagenomics

Conventional diagnostic techniques rely on culturing pathogens in selective media or detecting pathogen-associated biomarkers, and are therefore effective at identifying infections where suitable diagnostics are available. However, infections and their aetiologies are often complex, polymicrobial or of unknown origin[226,227]. Clinical metagenomics can offer a more comprehensive solution to traditional methods by effectively unveiling the complex microbial composition of the human microbiome – particularly valuable for polymicrobial infections that are often severe and challenging to detect and treat using conventional methods[228–230].

Metagenomic data can be used for monitoring antibiotic resistance[231,232], identification and surveillance of pathogenic strains[233], tracking in-host microbiome evolution[234], biomarker discovery[235] and detecting virulence factors and toxins of different environments in patients[236]. Information gathered from these strategies holds incredible potential for personalized medicine[137,236–238]. For example, clinicians can tailor treatment strategies more precisely by analysing the resistome of a clinical sample, reducing the risk of promoting antibiotic resistance[239]. Similarly, the detection of genes encoding toxins such as colibactin in complex biological samples from patient samples can serve as an early indicator of disease risk, such as colorectal cancer[240,241]. By integrating such detailed diagnostic information with insights into the complex microbial ecosystem of each infection, clinicians can more accurately predict clinical outcomes and tailor treatment strategies to individual needs[228,229,242–245].

Although still in its infancy, metagenomics has proven to be highly effective in early pathogen diagnosis, offering a short detection time and a high sensitivity[246]. It shows improved clinical outcomes in diagnosing culture-negative sepsis and infections that are difficult to diagnose, such as pneumonia, severe diarrhoea and meningitis[247–250]. Clinical metagenomics has emerged as an important tool in understanding the associations between microbial composition and various pathologies, such as cystic fibrosis, inflammatory bowel disease and colorectal cancer[251–253]. Additionally, metagenomic strain profiling workflows can improve clinical diagnostics by providing tools to track strain-level variability within complex metagenomic datasets[239,254]. For example, metagenomic strain profiling is invaluable for monitoring engraftment and tailoring treatment strategies to individual needs in the context of faecal microbiota transplant[239]. Furthermore, metagenomics has a crucial role in identifying emerging and novel pathogens, a capability that has become increasingly important following the COVID-19 pandemic[255–257].

Wide-scale implementation of metagenomics into clinical practice has lagged behind basic research applications, largely due to challenges associated with standardization, reproducibility, cost, slow turnaround time and regulations[258]. However, the declining costs, advancements of NGS technologies and the emergence of innovative computational tools can pave the way for advancing the field of clinical metagenomics. Standardization and validation of practices across clinical centres continues to present a significant challenge. In particular, standard operating procedures, the choice of DNA extraction methods, sample handling and computational analyses require high levels of standardization in clinical metagenomics. Moreover, divergent practices and protocols can adversely impact the interpretation of clinical metagenomic data influencing subsequent diagnostics.

Tracking the spread of disease and surveillance of pathogens

Traditional pathogen surveillance method strategies relying on cultured isolates face challenges related to resource demands, including the need for extensive laboratory materials, infrastructure, skilled personnel and the time to obtain pure isolates. Metagenomics offers a promising alternative to these methods by enabling culture-free detection of known and previously undetected viruses, bacteria and fungi, and can also identify virulence and resistance determinants in microbiome samples[257]. As metagenomics techniques continue to advance, they hold the potential to identify and track threats to the health of humans and farm animals as part of a comprehensive One Health approach[258].

Strain tracking, which involves the continuous monitoring of microbial strains over time and at different locations, has provided evidence of vertical transmission of microorganisms from maternal breast milk to the infant gut microbiome[259]. Strain tracking holds important clinical applications in the context of outbreaks of antibiotic-resistant pathogens in hospital settings. Metagenomics and other NGS approaches such as targeted whole-genome sequencing can be used to trace the spread of individual strains between patients, allowing healthcare facilities to respond appropriately with isolation measures and specific antibiotic interventions[260]. Bioinformatic pipelines, such as MIDAS[261], StrainPhlAn[262] and SameStr[263], have been developed specifically for strain tracking and population genomics across multiple samples. These tools are tailored for the taxonomic classification of reads from similar genomes and for detecting microorganisms of lower abundance where traditional taxonomic assignment methods might fail to differentiate sequences at the strain level.

Metagenomics can be used to monitor microbial communities for the presence of known and emerging pathogens by systematically analysing samples from diverse environments such as water, soil, sewage and air, facilitating the establishment of baseline microbial diversity data that can serve as reference for the surveillance of microbial temporal trends. Notable examples and initiatives of pathogen surveillance are discussed in Supplementary Box 8. Eventually, metagenomics could be used for developing and implementing metagenomic-backed policies and acting as an early warning system for the emergence of new pathogens or antimicrobial resistance for improved public health outcomes[56,264,265].

Environmental health monitoring and conservation

Metagenomics can offer powerful tools for identifying microbial communities involved in processes such as nutrient cycling[266,267], pollutant degradation and bioremediation[268]. Soil microbiomes, comprising bacteria and fungi, facilitate nutrient cycling by decomposing organic

matter and releasing essential nutrients such as nitrogen and phosphorus, with diazotrophic bacteria (such *Rhizobium*) enhancing soil fertility by converting atmospheric nitrogen into bioavailable ammonia[269]. Additionally, specific bacteria release organic acids, solubilizing phosphorus compounds to aid plant uptake, a critical process for agricultural productivity and soil health[270]. In pollutant degradation, bacteria such as *Pseudomonas* secrete extracellular enzymes and glycoconjugates that enhance the breakdown of organic pollutants, metabolizing hydrocarbons in contaminated soils into simpler, non-toxic compounds through oxidation, a process essential for mitigating the long-term accumulation of industrial pollutants[271]. Marine microalgae, such as *Chlorella* and *Spirulina*, have a significant role in bioremediation by biosorbing heavy metals and pollutants from wastewater, offering a sustainable solution for environmental clean-up[272,273]. Identifying microbial communities capable of degrading pollutants and elucidating the key degradation pathways would assist in the development of tailored bioremediation strategies to enhance our ability to remove a diverse array of environmental contaminants and perform real-time monitoring of microbial community dynamics during bioremediation processes[274,275]. Metagenomics can also be used to assess the health and biodiversity of ecosystems[276,277], for example by monitoring microbial community changes to determine the impacts of human activity, climate change and other environmental factors[278]. Microbial communities have a crucial role in supporting the health and resilience of animals, insects and plants, and metagenomics may be applied to support the conservation of endangered species[279].

Environmental metagenomics can be used for disease surveillance and outbreak prediction by monitoring microbial diversity in places such as hospitals, water plants and farms, allowing for early intervention[257]. It could also be used for to the discovery of new antibiotics from diverse environments, which is crucial for battling antibiotic-resistant bacteria[280,281]. Metagenomics studies of built environments can provide insights into indoor air quality and pathogen dynamics[282].

Sustainable agriculture

Understanding the soil microbiome is essential for sustainable agriculture and managing soil health[283]. Potential applications of metagenomics in this space could include investigating how plants interact with soil microorganisms to design microbial-based solutions for enhancing crop productivity, improving nutrient cycling and boosting disease resistance[284,285]. Metagenomics can also aid in the identification of soil microorganisms that are beneficial to crop species, in order to facilitate the development of biofertilizers and microbial inoculants and reduce chemical fertilizer use[286,287]. Soil metagenomics has been invaluable for studying the rhizosphere microbiome — the microbial community surrounding plant roots — helping pinpoint beneficial microorganisms that suppress plant diseases to promote sustainable agricultural practices. For more about soil metagenomics and metagenomics in agriculture, see Box 1.

Reproducibility and data deposition

Following the first commercial high-throughput sequencer introduced in 2005 (ref. 288), the cost of high-throughput sequencing has greatly decreased, leading to a rapid increase in the amount of metagenomic data that are being generated and stored in public repositories[289] (Table 1 and Supplementary Table 4). Such repositories serve as a vital resource by facilitating the dissemination of a diverse array of publicly available data to the wider scientific community, including metagenomic sequencing data, sequence annotations, geospatial data and computational resources (see more about public repositories for raw metagenomics data in Supplementary Box 9). The reliability and robustness of metagenomic analyses fundamentally depend on the completeness and accuracy of reference sequences from genome databases (Table 2). Existing databases cover a range of different taxa and exhibit varying levels of completeness and annotation quality in genome assemblies.

For new users of metagenomics techniques, understanding what metadata need to be considered is crucial for ensuring comprehensive and standardized reuse of existing data from public repositories. Metadata include varied information such as sequencing parameters, sample and collection, and quality control measures[290,291], which are summarized in Supplementary Table 5.

The Genomic Standards Consortium[292], BioProject and BioSample project[293] are collectively focused on establishing standardized protocols and minimum information standards, streamlined data organization and enhanced metadata quality. By promoting uniform data organization and improving metadata quality, these groups aim to ensure that datasets are more accessible and comparable across studies.

Microbial reference genome resources

Microbial reference databases might include both complete and draft genomes[78]. However, specialized reference genome databases often implement rigorous quality control procedures to ensure that only assemblies meeting high standards of sequence and annotation quality are included. For example, the RefSeq database excludes assemblies derived from environmental samples owing to concerns about the accuracy of organism assignment and potential cross-contamination. Differences between reference sequence sources across databases are due to minor genomic variations among organisms and their genome organization, potentially leading to inconsistencies. Additionally, different databases can show conflicting taxonomic labels for identical species or strains[15]. These and other discrepancies can exist among commonly used reference sequence databases such as RefSeq[294], Ensembl[295] and the Pathosystems Resource Integration Center (PATRIC)[296]. Some species are uniquely represented in only one database, leading to a difference in genomic quality that cannot be reconciled by using another[15]. Inconsistencies and discrepancies between sequence databases can affect the outcomes of bioinformatic analyses, limiting both the accuracy and reproducibility of metagenomic studies[15]. Researchers are therefore urged to meticulously evaluate both the quality of raw metagenomics data and the accompanying metadata to guarantee robustness and reliability in their analyses. We suggest users consider the latest and most complete databases, such as those with regular updates such as RefSeq or the Genome Taxonomy Database (GTDB), remain consistent with data analysis and avoid switching databases unless for a good reason; and remain cautious when comparing results generated from distinct database by double-checking differences in genomic contents. A master database incorporating multiple resources, which is a complex and time-consuming task, would alleviate these issues.

The frequency and rigour of updates vary greatly across different platforms. GenBank[297], for example, relies on original submitters to update their sequence submissions. GenBank also has less strict curation policies than other databases, primarily focusing on adding

Box 1 | Environmental metagenomics applications

Food safety metagenomics

Metagenomic studies have profiled the microbial communities present across various stages of the food production and supply chain industries. Samples include food products, ingredients and environmental samples such as processing plant surfaces[381], storage temperatures[382], packaging[383], irrigation water[384] and even additives such as salt[385]. Leveraging metagenomic techniques can advance our understanding of microbial interactions within the food industry and enable more refined protocols for the early detection and prevention of food-borne pathogens[386].

Soil metagenomics

Metagenomics research has illuminated microbial diversity across various soil types and highlighted dramatic alterations in soil microbiome composition over time due to industrialization and increasing contamination. For example, metagenomics analyses have revealed the complex interactions between pesticides and microbial genes involved in pesticide degradation, underscoring the potential of soil microbial communities to mitigate pesticide contaminations[387]. They have also aided in discovering beneficial microorganisms that promote plant growth and protect against diseases, contributing to more resilient agricultural systems[284,388]. In addition to soil-based studies, metagenomic research has been applied to farm animals to track viral pathogens and determine the relationship between the microbial community and host nutrition and metabolism[389].

Water metagenomics

Water ecosystems including oceans[45], lakes[390–392] (including artificial[393] and volcanic lakes[394]), rivers[395] and mangroves[396] have been extensively studied through metagenomics. Analyses of deep ocean sources using functional metagenomics have revealed a surprising diversity of metabolic strategies among microorganisms[45]. Results from metagenomic analyses of wastewater and contaminated groundwater provide crucial insights for developing effective recycling and bioremediation methods[397]. Metagenomic tools have also enhanced the assessment of microbial communities in drinking water for water quality monitoring[398]. These methods can track indicators of faecal pollution, detect common bacteria in polluted water and explore the presence of viruses.

Air metagenomics

Metagenomic studies of air samples are challenging due to low microbial density and the lack of standardized methodologies[344]. However, these studies revealed a positive correlation between certain airborne microbial genera and mortality rates in patients with respiratory diseases and identified positive and negative correlations between anthropogenic activities and airborne microbial communities[399]. Hospital air directly impacts patient health and is a crucial focus for metagenomic research[400–402]; shotgun metagenomic approaches have identified an abundance of opportunistic pathogens in hospital air[264,401].

Metagenomic bioengineering

The integration of synthetic biology with functional metagenomics has revolutionized the discovery and exploitation of novel genes and biochemical structures from metagenomic profiles. This synergy has led to the identification of unique enzymes, and molecular interactions with therapeutic potential. For example, the interaction of *N*-acyl-amide synthases and G-protein-coupled receptors[403] and the discovery of sialidases within the glycoside hydrolase family (GH156)[404] in gastrointestinal bacteria underscore the potential of metagenomic bioengineering to reveal previously inaccessible molecular targets. The application of synthetic biology on marine metagenomics has enabled the reconstruction of complete genomes of previously unknown microorganisms, leading to the discovery of new genes and metabolic pathways with significant biotechnological potential[405]. Further, metagenomics serves as a valuable resource for mining new enzymes from microbial community metabolomes through functional screening and sequence-based approaches[406]. A wide range of enzymes, including lipases, cellulases and proteases, have been discovered, with metagenomic analyses promising even more discoveries in the future[407].

Non-human microbiome

Profiling non-human species has been essential in comparative metagenomics. For example, a metagenomic study on the gut microbiomes of various non-human primate cohorts revealed an array of previously unidentified microorganisms and a 20% shared composition of species with the human microbiome[408]. Further, the use of metagenomic data from non-human species can shed light on how microbial communities co-evolve between and within species. A metagenomic study across three species of carpenter bees evaluating the core microbiome among these species revealed genomic variations[409].

new uploads rather than curating existing data. By contrast, the RefSeq database[298] is regularly updated to refine annotations and integrate new information, adhering to specific quality standards[298].

The scope of organisms present within reference databases varies widely. Some databases, such as RefSeq, cover a broad range of organisms across all living kingdoms, including bacteria, viruses, archaea and eukaryotes. By contrast, other databases narrow their focus to specific groups of species; for example, FungiDB[299] and the Joint Genome Institute (JGI) MycoCosm[300,301] specialize in fungal reference sequences, and the PATRIC database[296] (part of BV-BRC[302]) is dedicated to bacterial and viral genomes.

The usability of databases for downloading reference genomes can vary. Most databases provide easy access and bulk download to all their stored genomes and other related data through public access file transfer protocol (FTP) (Supplementary Table 4). However, others including the JGI MycoCosm[300,301], Microbial Genome Database (MBGD)[303] and Human Reference Gut Microbiome (HRGM)[304] require a user registration step.

The continuous refinement and expansion of reference genome databases has a crucial role in advancing the field of metagenomics and its applications in understanding microbial communities. Despite ongoing efforts to compile comprehensive reference sequence

Table 1 | Most-used publicly available repositories for metagenomic data

Database	Total metagenomic samples (approximate)	Description
ENA/SRA	2,000,000	Largest public repository of high-throughput metagenomic data
MGnfy	343,000	Platform for harmonized assembly, functional and taxonomic analysis of diverse meta-datasets including sample storage and inter-study associations
MG-RAST[a]	512,000	Upon upload, the pipeline performs quality control, functional and phylogenetic analysis; real and simulated data can be stored as public or private
Tara Oceans	35,000	Database focusing on sunlit ocean life, consisting of seawater and plankton samples from 210 ocean stations; includes data from the Global Ocean Sampling expedition, Pacific Ocean Virome project, National Center for Biotechnology Information (NCBI) reference genomes and Moore Microbial Genome Sequencing Project

Sample number is estimated without counting amplicons. [a]MG-RAST is currently not actively supported by any organization.

databases, a large portion of metagenomic reads in a sample remain unexplained, failing to be assigned to any known taxa[6,305,306].

Reproducibility and reusability of public metagenomic data

High-quality public metagenomic datasets can facilitate the reuse of previously published metagenomic data, supporting open science principles that promote transparency and collaboration within the research community[307]. Easy-to-use formats and complete metadata are essential for maximizing the usefulness of these datasets. However, concerns regarding data quality, metadata incompleteness and the heavy computational requirements of analysing raw metagenomic data have resulted in the under-use of available datasets by the scientific community[308,309].

One way to reuse metagenomic data is through meta-analyses[290,310]. The aggregation of multiple studies into large meta-analyses has become a powerful practice for uncovering novel insights across diverse domains such as human health[311], microbial ecology[312], virology[313] and the identification of promising drug targets[314,315]. Effectively performing meta-analyses of existing metagenomic studies is subject to numerous challenges associated with data reuse, such as disparities in data generation, metadata incompleteness and limited public access to relevant raw data[290].

Some recent cloud-based repositories provide integrated computational resources and workflows to facilitate the analysis of large datasets. Platforms such as MGnify[316] offer automatic analyses through established pipelines and provide processed results for immediate research use. Similarly, the NCBI Cloud Resources and the Galaxy Project provide cloud-based access to computational tools and databases, enabling complex bioinformatic tasks without the need for high-end local infrastructure. The Galaxy Project further simplifies the process by offering a user-friendly, web-based interface that does not require programming skills. CyVerse[317] provides integrated cyberinfrastructure to support data storage and processing across both public and private clouds. Cloud-based repositories will be crucial for democratizing access to metagenomic data analysis, allowing more researchers to leverage advanced computational tools and infrastructure. However, scaling such resources to handle efficient analyses of thousands of samples remains a challenge. Further, the associated costs of these resources could be burdensome, particularly for researchers from low-resource universities and countries.

Limitations and optimizations

Several factors must be considered to counteract systemic biases and mitigate confounding factors within the study design phase of metagenomic analyses. This section outlines specific limitations and confounding factors that can be present in metagenomics analysis and discusses strategies for how they might be addressed.

Experimental and sampling biases

Sampling from different geographic locations or under varying environmental conditions can confound results due to differences in microbial populations[318]. Environmental variables such as pH, temperature, moisture and nutrient availability have important roles in shaping microbial community structures. Further, differences in the way samples are collected[319], the duration of sample storage[320] and the storage medium used[321] can further contribute to experimental biases. Failing to consider and properly curate the metadata might distort the relationship between microbial composition and the variables of interest.

Behavioural aspects can exert considerable influence on the microbial composition of human or animal microbiomes. For example, dietary habits have been shown to shape the dynamics of the human gut microbiome, with disruption of the microbiome linked to chronic conditions such as cardiometabolic diseases and type 2 diabetes[156,322,323]. Failure to control for behavioural factors such as diet and lifestyle could confound study results[324]. Host genetics can also affect the diversity and abundance of specific taxa in the host-associated microbiome[325,326]. Failure to account for genetic associations with the microbiome could confound the interpretation of results, obscuring the true reciprocal relationship between hosts and their associated microbiota. Metagenome-wide association studies have been used to account for genetic associations; for example, applying metagenome-wide association studies on the tongue dorsum and saliva has revealed five genetic loci associations with the oral microbiome[327].

In longitudinal studies, samples analysed after extended storage periods might differ from those sequenced closer to their sampling dates and researchers must cautiously interpret minor but consistent declines in microbial richness over time[320]. Adherence to rigorously standardized protocols is important to mitigate sampling and storage biases and ensure consistency and reliability of the data. Strategies to address storage issues include collecting biological and technical replicates, using standardized collection devices and procedures, and maintaining aseptic conditions throughout storage and analysis to prevent contamination. Rapid freezing of samples at −80 °C is also recommended for optimal preservation of sample integrity. When freezing is not feasible, the consistent use of preservatives is crucial. The duration of storage must be documented and considered during analysis to ensure accurate interpretation of results.

Three primary sources can introduce bias in metagenomic analyses during sample preparation: variable DNA extraction, contaminant

DNA introduction and inclusion of DNA from non-viable host cells or microorganisms. The use of different DNA extraction methods can result in microbial community profiles that are distinct from those present in the original samples. For example, some microorganisms with resilient cell walls are inadequately represented in some metagenomic datasets owing to inefficient cell lysis. Moreover, different extraction methods might yield varying amounts of DNA or exhibit biases towards certain types of microorganisms. Consequently, DNA extraction represents one of the most biased steps in metagenomic data generation[328,329]. When managing DNA extraction biases, it is essential to employ validated kits or protocols while ensuring consistency across samples by following a standardized protocol. Documentation of extraction batches and inclusion of this information as a covariate in downstream analysis are also important for mitigating batch effects.

Contamination from the sampling and laboratory environments, equipment, reagents and even the researchers themselves can introduce foreign DNA sequences into metagenomic data, complicating the interpretation of results[330]. Addressing contamination requires the implementation of rigorous quality control measures to identify potential sources of contamination and minimize its effects. Contamination from external sources might originate from kits and reagents that commonly contain small amounts of contaminating bacteria, which can vary across manufacturers and facilities[330,331]. Additionally, cross-contamination can occur due to the inadvertent transfer of sample DNA, indexes (barcodes) and amplicons between samples, further exacerbating this issue. Although low-biomass samples are more susceptible to the deleterious effects of contaminants, any sample can be affected[330–332]. Implementing appropriate negative controls such as sampling blanks, DNA extraction blanks and no-template amplification is recommended to address metagenomic contamination, particularly for low-biomass samples. Through such measures, contaminants can be distinguished from endogenous (native) taxa and subsequently removed from the data[333,334]. The RIDE checklist offers valuable recommendations to reduce metagenomic contamination[335], emphasizing the importance of reporting experimental design, incorporating negative controls, determining the level of contamination and evaluating the influence that contaminant taxa have on the interpretation of metagenomic data[335]. Contamination can be further addressed computationally via algorithmic models that are able to statistically detect metagenomic contamination via analysis of the read frequency[333].

Some metagenomic samples contain high levels of host DNA with low microbial biomass. Such samples include human and animal milk, which can contain up to 95% host DNA[32,336], host tissues, saliva and other bodily fluids. An abundance of host DNA presents challenges for distinguishing between host and microbial reads, particularly for novel microorganisms. Notably, the plant holobiont is susceptible to contamination from sources such as soil microorganisms, co-amplified plant organelles and human DNA. These contaminants necessitate rigorous sampling protocols, specialized preparatory steps and robust decontamination during bioinformatics analysis. Plant sample preparation should involve washing with sterilized solutes to remove potential contaminants; sequencing these reagents separately helps identify contaminants, preventing their interference with microbiome analyses and ensuring the detection of low-abundance community members[329,337]. Common sources of host contamination in plant metagenomics include co-extraction of chloroplast and mitochondrial DNA during milling and physicochemical lysis, as well as human DNA and relic DNA from the rhizosphere. These contaminants can obscure estimates of soil microbial diversity, impacting the analysis of root samples (rhizosphere soil) and other plant tissues[338–340]. Various strategies have been developed to address the challenges presented by host DNA, including selective extraction to remove host cells[341,342], post-extraction methods that enrich for microbial genomes[32,343] and bioinformatic approaches to filter out host-associated reads post sequencing. It is important to note that host depletion approaches might introduce bias in the sequencing of specific elements within a metagenome[336,342].

Low microbial concentrations. Microbial concentrations can vary across metagenomic samples, posing a challenge for metagenomic research. This issue is particularly relevant in clinical metagenomics

Table 2 | Examples of publicly available databases that store microbial reference genomes

Database	Total species (approximately)	Contents
RefSeq	125,000	Comprehensive, non-redundant, well-annotated set of reference genomic sequences presented in the GenBank database, including archaea, bacteria, fungi and viruses
Ensembl[377]	300,000	Vertebrate genomes, plants, fungi, bacteria and protists
VEuPathDB[378]	500	Eukaryotic pathogens (protists and fungi) and relevant free-living or non-pathogenic species
BV-BRC[a] (ref. 302)	87,000	Combines the Pathosystems Resource Integration Center (PATRIC), the Influenza Research Database (IRD) and the Virus Pathogen Database and Analysis Resource (ViPR), and includes bacterial and viral pathogens and archaeal genomes
Unified Human Gastrointestinal Genome (UHGG)[379]	204,000	Non-redundant genomes for human gut prokaryotes
Joint Genome Institute (JGI) 1000 fungal genomes project (JGI 1K)[a] (ref. 300)	388	Fungal and algal genomes
Human Reference Gut Microbiome (HRGM)[a] (ref. 304)	232,000	Non-redundant genomes for representative prokaryotic species
Microbial Genome Database (MBGD)[a] (ref. 380)	15,000	Incorporates all complete genome sequences of bacteria, archaea and unicellular eukaryotes including fungi and protozoa available at the National Center for Biotechnology Information (NCBI) genome FTP site

Links to access databases and file transfer protocol (FTP) servers are provided in Supplementary Table 4. [a]Hyperlink for database does not link to an FTP server.

as some body sites, such as urine[342], skin[31], the lungs[343] and blood[328], contain low microbial concentrations. Therefore, the volume necessary for obtaining sufficient microbial biomass from non-stool samples should be considered[344,345]. For example, a study on DNA yield from different environmental samples (soil, water and air) revealed that air samples require a much larger volume of sample to yield 5 ng of DNA when compared with soil and water[344]. Alternative approaches have been presented to study the microbiomes of ultra-low-biomass surfaces and the use of negative controls should be able to detect genomic contamination that can result from the use of DNA-based reagents[346]. If appropriate measures (such as sample volume[344], reagent use[330,346] and the incorporation of negative controls[346]) are not taken, metagenomic analysis of low-biomass samples can lead to biased interpretations of study results.

The limit of detection for sequencing-based assays is an important consideration when using metagenomic analysis tools for clinical applications, even in samples with high bulk microbial concentrations. For example, *Clostridioides difficile* is an important human pathogen that has a relative abundance of less than 3% of the total microbial populations in stool of patients with active *C. difficile* infection, making diagnosis challenging[347]. Clinical diagnostic protocols rely on enrichment techniques, such as ex vivo culture or PCR, or focus on highly abundant expressed genes known in *C. difficile* infection[347]. Untargeted metagenomic sequencing would require extreme deep sequencing efforts to capture hard-to-detect pathogens such as *C. difficile*. Achieving clinically meaningful interpretations of metagenomic data is essential for high fidelity and requires high-quality sequences that accurately represent the microbial community present in the sample. The accuracy from high-fidelity data ensures the ability to precisely predict antibiotic susceptibility. For example, unreliable data can lead to misidentification of resistance genes compromising treatment options. Therefore, the purpose in refining the sequencing process must mitigate errors and ensure robust data to support clinical decisions that are reliable. Although sequencing costs are declining, the substantial increase in the cost of the computational analyses needed to process, store and extract the sequences belonging to the targeted pathogen may be overlooked. Therefore, it is likely that clinical applications of metagenomics will continue to require the use of target-enrichment protocols in order to have a meaningful impact in clinical practice.

Studies of the placental microbiome are a powerful example of how low-biomass samples can affect results from metagenomics studies. Studies have suggested that the placenta has a distinguishable microbiome[348–350]. However, the placenta has also been described as a nearly sterile organ with a low biomass concentration, and samples extracted from the basal plate can easily be influenced by contaminants[330] from reagents[330,351] or the surrounding environment, such as the maternal skin[352], which has led to speculation regarding whether it possesses a unique microbiome. Ultimately, studies have presented evidence that what was previously interpreted as a placental microbiome was, instead, likely contamination[353–356].

Several optimizations exist to improve the resolution of low-biomass microbial communities and minimize contamination in these samples, such as the use of sterile protocols and appropriate negative controls, as outlined above. Given that reagents can be a source of contamination from either manufacturing or laboratory contamination, these should be treated with DNase as a preliminary step to reduce contamination of low-biomass DNA samples prior to DNA amplification[351]. In the case of the placenta, noted above, specific sterile procedures might include performing thorough disinfection of the maternal skin and using a sterile drape to cover the patient from the armpits to the knees before and after caesarean[352].

Constraints in metagenomic assembly. Challenges in de novo metagenome assembly often arise because of the sheer complexity of microbial communities. Metagenomic samples frequently contain a mixture of microorganisms with different abundances and genomic compositions, including microorganisms that are genomically difficult to distinguish. Such samples might contain microbial genomes with vastly different coverage depth[357] and a complete assembly might only be possible for a few high-abundance organisms.

A high-quality metagenomic assembly is dependent on several features of the sequencing platforms and associated protocols that are used, such as read length, throughput and error rate. At low coverage depths (typically below threefold to fivefold), many segments of microbial genomes are not sampled by reads owing to the random nature of the sequencing process. Consequently, the assembly software can only reconstruct a highly fragmented approximation of microbial genomes, with sequence continuity disrupted in regions where coverage is low or absent.

Read length impacts the quality of the sequence assembly owing to the ambiguity caused by repetitive sequences. The presence of repeats hinders the reconstruction of microbial genomes as the assembler might not be able to determine the correct genomic placement of reads contained within repetitive regions. This issue can be addressed by using longer read lengths, which reduce the probability that reads lie completely within or greatly overlap repeat regions[358]. Paired-end reads are also often preferred over single-end reads as they can help bridge gaps in assembly (a process called scaffolding) or resolve ambiguity introduced by repeats[359].

Metagenomic assemblers must contend with the uneven depth of coverage of contigs within samples (an artefact due to the varied abundances of organisms in a sample), a factor that complicates the detection of repeats and is further confounded by segments of DNA that are shared between different organisms (instead of just being repeated within a single genome)[113,117]. To account for the effects of sequencing errors and strain-level variation without exacerbating the impact of repeats, metagenomic assemblers tend to integrate assembly graphs generated with multiple *k*-mer sizes. Shorter *k*-mers are more able to overcome sequencing errors or strain variation, whereas longer *k*-mers are more effective at resolving repeats. Some assemblers (such as metaSPAdes[117]) aggressively attempt to 'smooth' out sequencing errors and strains in order to generate longer contiguous segments (contigs), whereas others (such as MEGAHIT[114]) are more conservative, resulting in more fragmented assemblies.

Long-read data can overcome some of the challenges posed by repeats and even strain variants, although they are associated with higher error rates than short-read data. Similar to short-read metagenomic assemblers, long-read metagenomic assemblers are often developed upon tools initially used to assemble single genomes. Hi-C technology has also been successfully adapted for metagenomics owing to its ability to link together genomic contigs from individual organisms, helping mitigate the issue of genomic fragmentation typically seen in metagenomic data[86,360–363].

Challenges in taxonomic and functional analysis. Challenges associated with taxonomic and functional profiling from metagenomic data include issues regarding accurate classification amid database biases, as well as the complexities of predicting functions and metabolic

Table 3 | Details of publicly available initiatives dedicated to the standardization of metagenomic data

Initiative	Total metagenomic samples (approximate)	Description
HumanMetagenomeDB 1.0	69,000	Standardizes human metagenomic data; sequencing data are not available on the website but a script is provided for downloads
TerrestrialMetagenomeDB 2.0	20,000	Focused on the standardization of metadata for terrestrial metagenomes; similar to HumanMetagenomeDB 1.0, developers offer a download script
PlanetMicrobe	2,000	Specializes in the standardization of marine or non-marine aquatic samples
MetaSUB Consortium	20,000	Standardization of collection, storage, transport, processing and analysis of surface, air, water and sewage samples from urban and rural areas
Serratus Database	5,700,000	Access to raw virome sequencing data including RefSeq vertebrate viruses, GenBank, Coronaviridae and full-length RNA-dependent RNA polymerase sequences, standardizing the processing of contigs including assembly and annotation
National Microbiome Data Consortium (NMDC)	2,900	Integrates the standards of multiple organizations to evaluate quality and access for metagenomic data, in addition to other multi-omics data such as proteomics, metatranscriptomics and metabolomics
Critical Assessment of Metagenome Interpretation (CAMI)	526	Comprehensive evaluation of metagenomic tools, offering reliable performance data to guide accurate interpretation and method improvement

interactions within diverse microbial communities. Overcoming these hurdles will require interdisciplinary collaboration, advancing computational tools and refining analytical methods.

Taxonomic profiling is inherently a reference-based analysis, where the accuracy of a predicted profile depends on the quality and comprehensiveness of the reference database used. Incorrect selection of reference genomes can result in the exclusion of entire taxonomic kingdoms[364] and incorrect taxonomic assignment[15]. Differences in nomenclature between taxonomic databases can also alter the distribution of reported taxa. For example, the International Code of Nomenclature of Prokaryotes (ICNP)[365] nomenclature integrates genomic information and observed physical characteristics of microorganisms, whereas the GTDB[366] nomenclature is exclusively reliant on genome phylogeny[367]. Additionally, it is important to acknowledge that reference databases are biased towards extensively researched organisms[368]. Horizontal gene transfer can further blur taxonomic profiling as genes move between microbial genomes. In Supplementary Box 10, we provide an example of how genome database-related biases affect metagenomic analysis, and the transition from taxonomic to functional profiling to improve strain-level resolution.

Outlook

The complexity, multidisciplinary nature, and conceptual and technical challenges of metagenomic research have led to the need for large-scale metagenomics initiatives (Table 3 and Supplementary Table 3). Such efforts were established for setting advanced standards in data collection, storage and sharing, while also assessing the impact of these initiatives on principles and values such as trust, confidentiality and privacy. Large-scale metagenomic initiatives enable the creation of community data resources and common standards while fostering technological advancements, including the development of new tools and shared data and software resources. Consequently, large-scale collaborative metagenomic initiatives have been critical for enhancing our understanding of global biological diversity of environments and human health. Examples of such initiatives include those focused on unmanaged landscapes and aquatic environments, such as Tara Oceans[4] and the Global Ocean Sampling Expedition[369], which explored the biological diversity in seawater and sediments. In parallel, the TerraGenome[370] project explored microbial diversity in soil. Other initiatives have focused on characterizing managed ecosystems and the microbial communities within urban environments, including the Earth Microbiome Project[7] and the MetaSUB initiative[371]. Another large area of research is host-associated habitats, where initiatives such as the Human Microbiome Project[372], the BeeBiome Consortium[373] focus on examining microbial interactions with hosts including the influence on host health and their interactions within ecosystems. These projects were characterized by centralized management and meticulous design to ensure the production of high-quality data and rigorous downstream data analysis.

We expect that global metagenomic initiatives will continue to support the establishment of broad foundational principles and innovative technologies and methods, given that such advances are more effectively implemented within the context of multi-institutional and highly replicated studies than in traditional single-investigator projects. Additionally, such large projects provide unique opportunities for public outreach and educational development within the metagenomics field. We note, however, that implementing and executing international and multidisciplinary metagenomics projects demands a high level of collaboration and coordination that extends beyond the possibilities of what individual research groups can typically manage. In particular, owing to their international character reach, broad scope and the involvement of numerous investigators, these initiatives necessitate carefully developed management plans, along with funding that is specifically allocated to enhance communication and promote effective collaboration. Despite these challenges, international and multidisciplinary projects offer an excellent opportunity to train a new generation of young scientists, equipping them with skills necessary for collaborative and large-scale science (Supplementary Box 11).

Optimal insights are often achieved through concerted multi-investigator and multidisciplinary efforts by integrating various methods, such as sequencing, functional-expression analysis, metabolomics analysis and deep phylogenetic analysis. Consequently,

Glossary

Alpha diversity indices
Indices that measure species diversity within a single, local microbial community or sample and consider both the number of different species (richness) and evenness of their distribution, as assessed through indices such as the Shannon index or Simpson's index.

Amplicon sequencing
A procedure in which scientists target, amplify and sequence marker genes, typically the 16S and/or 18S ribosomal RNA genes.

Antibiotic resistance genes
Specific genetic sequences within microbial communities that encode factors conferring resistance to antibiotics; these genes are often located on plasmids or transposons and can be transferred from cell to cell by conjugation, transformation or transduction.

Antimicrobial resistance
The developed capability of bacteria and fungi to resist drugs designed to kill them.

Contigs
Contiguous sequences of DNA that are assembled from overlapping sequence reads.

de Bruijn graph
A data structure used in genomics to represent overlaps between sequences, where nodes represent *k*-mers (substrings of length *k*) and directed edges represent overlaps of *k*−1 bases between consecutive *k*-mers; the sequence of the genome being assembled can be 'read' by traversing the graph and concatenating the sequence of the *k*-mers encountered.

Environmental DNA
Genetic material obtained directly from environmental samples such as soil, water and air.

Environmental metagenomics
The study of the microbial community present in a natural ecosystem such as water, soil or air.

Functional metagenomics
The study of the genomic roles and interactions within a microbial community; this includes tasks such as gene prediction, functional annotation and functional profiling to characterize its genomic composition and metabolic and enzymatic activities.

Functional potential
The presence of genomic elements (such as genes) from which the physiological processes of a microorganism can be inferred, but not verified without direct experimental evidence.

Functional profiles
Characterizations of the potential biological functions and metabolic pathways within a microbial community.

High-performance computing clusters
A network of computers that work in tandem to efficiently tackle intensive computational tasks.

Horizontal gene transfer
The exchange of genetic material between organisms that are not in a parent–offspring relationship.

k-mers
Substrings, of length *k*, derived from a longer DNA or RNA sequence, used in bioinformatics for tasks such as assembling genomes, analysing sequence composition and identifying sequence similarities.

Marker genes
Evolutionarily conserved genes with one or more variable regions that are used as an evolutionary clock to delineate phylogenetic lineages and classify microorganisms into taxa.

Metabolomics
Experimental and computational approaches used to characterize the metabolite profiles found in microbial communities.

Metagenome-assembled genomes
(MAGs). Genomes or collections of genome fragments originating from a single organism extracted from a microbial community.

Metagenome-wide association studies
Statistical metagenome-wide studies that involve the identification and association of genetic loci and genomic information with disease in relation to the host and its environment.

Metagenomic assembly
The process of reconstructing individual genomes or genome fragments (contigs) from the sampled DNA of a microbial community.

Metagenomic binning
The grouping of contigs into discrete bins or collections that represent individual organisms or taxa.

Metagenomic profile
A comprehensive overview of collective microbial genetic material from a sample, providing insights into species composition, functional potential and relative abundance.

Microbiome
A microbial community sampled from an environment where different species of fungi, bacteria, archaea and viruses can be present; the definition includes microorganisms and their functions and interactions.

One Health approach
A metagenomic method that connects metagenomic information of patients, animals and the environment to encompass and characterize the interaction of microbial communities to aid in the clinical diagnosis and treatment of humans.

Open reading frames
(ORFs). Continuous sequences of codons in a genomic region, starting with a start codon and ending with a stop codon, that has the potential to be translated into a functional protein.

Resistome
The landscape of genes present in the microbiome that are resistant to antibiotic treatment; antibiotic-resistant genes can be acquired through horizontal gene transfers, which is an evolutionary event where genes can be moved and adopted among a microbial community.

Sequencing depth
The number of times that a particular DNA or RNA nucleotide is read during the sequencing process. The average depth of sequencing coverage can be defined as LN/G, where L is the read length, N is the number of reads and G is the haploid genome length.

Shotgun metagenomics
An approach that enables the analysis of an entire microbial community without targeting specific species by randomly fragmenting the genomic material present in a metagenomic sample and sequencing it.

Taxonomic profile
A list describing the taxa in a microbial community along with their relative abundances.

Viral quasispecies
A group of closely related viral variants resulting from high mutation rates during viral replication forming a dynamic population of genetically diverse but related sequences, often referred to as a 'cloud' of mutants.

Virulence factors
The molecules that assist microbial pathogens to colonize a host at the cellular level.

such initiatives will benefit from the broad expertise of professionals from numerous fields, including geneticists, microbiologists, physicians, bioinformaticians and computational scientists. By integrating diverse expertise and perspectives, researchers can tackle complex challenges more effectively to drive innovation in metagenomic research. Furthermore, leveraging emerging technologies such as machine learning holds promise for facilitating both data analysis and interpretation. These tools offer opportunities to extract valuable insights from vast amounts of microbiome data, accelerating discoveries and enabling development of personalized approaches to healthcare. Promoting open science principles and data sharing, such as through the establishment of accessible repositories and platforms for sharing data, methodologies and findings, can maximize the impact of collaborative efforts, fostering transparency, reproducibility and community engagement.

Building on data values and ethical principles, large-scale metagenomic projects also increase outreach by inviting scientists from outside the specific consortium to participate in data exploration and analysis. Notably, the Critical Assessment of Massive Data Analysis (CAMDA) challenges and conferences have fostered community-driven data analysis for biological Big Data for two decades[374]. Since 2016, CAMDA has partnered with the International MetaSUB Consortium, featuring extensive datasets from large-scale sampling of mass-transit systems and other public areas across the globe generated during City Sampling Day actions. Thus, the CAMDA community has had the opportunity to explore MetaSUB data in the context of microbiome-based sample origin prediction or resistome characterization/antimicrobial resistance prediction, resulting in more than 20 publications to date. Currently, CAMDA participants are exploring new metagenomics data analysis-related areas including gut microbiome-based health assessment.

A critical component of large-scale metagenomic research is the refinement of bioinformatic tools, which require rigorous benchmarking[374]. One practical approach for benchmarking metagenomics tools is through community-organized challenge-based assessments. A notable example is Critical Assessment of Metagenome Interpretation (CAMI)[375], offering challenges dedicated to benchmarking metagenomic tools designed for various tasks such as genome assembly, taxonomic profiling and binning. These challenges delve into analysing the accuracy, run time and memory usage of bioinformatics tools, providing invaluable insights into their performance. The results obtained from such challenges are instrumental in identifying current limitations of existing computational methods, thereby paving the way for advancements in computational metagenomics.

The range of applications for metagenomics is continually expanding, with profound implications for public health, environmental management and biodiversity conservation. For example, metagenomics applications have significant potential in the clinical setting, and their implementation is revolutionizing infectious disease diagnostics and pathogen surveillance[376]. That is, the increased use of metagenomics in a patient-care context will allow clinical microbiology to move past simple pathogen identification towards a holistic diagnosis that includes a systematic breakdown of antibiotic resistance, virulence factors, host metabolic determinants and ecological context. Moreover, sequencing and sequence databases can now be leveraged to learn more about each pathogen diagnosis in the context of the patient, the hospital and the broader community. This ability will allow us to better understand the specific context of each infection, thus facilitating personalized medicine while also allowing epidemiologists to track infection trends in real time. Lastly, metagenomics can broadly enhance our understanding of microbial contributions to biodiversity, help us preserve microbial diversity, and assist in identifying rare and unique microbial taxa that will be important for both bioprospecting and the development of new sustainable biotechnological applications to drive future scientific innovations.

Published online: 23 January 2025

References

1. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).
2. Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
3. Rondon, M. R. et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
4. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
5. Gevers, D. et al. The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol.* **10**, e1001377 (2012).
6. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* **184**, e17 (2021).
7. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
8. Schloss, P. D. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio* https://doi.org/10.1128/mbio.00525-18 (2018).
9. Vandeputte, D., Tito, R. Y., Vanleeuwen, R., Falony, G. & Raes, J. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* **41**, S154–S167 (2017).
10. Wu, W.-K. et al. Optimization of fecal sample processing for microbiome study—the journey from bathroom to bench. *J. Formos. Med. Assoc.* **118**, 545–555 (2019).
11. Kennedy, K. M. et al. Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. *Nature* **613**, 639–649 (2023).
12. Bei, Q. et al. Extreme summers impact cropland and grassland soil microbiomes. *ISME J.* **17**, 1589–1600 (2023).
13. Devkota, S. Prescription drugs obscure microbiome analyses. *Science* **351**, 452–453 (2016).
14. Lombard, N., Prestat, E., van Elsas, J. D. & Simonet, P. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol. Ecol.* **78**, 31–49 (2011).
15. Loeffler, C. et al. Improving the usability and comprehensiveness of microbial databases. *BMC Biol.* **18**, 37 (2020).
16. Jones, J., Reinke, S. N., Ali, A., Palmer, D. J. & Christophersen, C. T. Fecal sample collection methods and time of day impact microbiome composition and short chain fatty acid concentrations. *Sci. Rep.* **11**, 13964 (2021).
17. Nayfach, S. & Pollard, K. S. Toward accurate and quantitative comparative metagenomics. *Cell* **166**, 1103–1116 (2016).
18. Bartolomaeus, T. U. P. et al. Quantifying technical confounders in microbiome studies. *Cardiovasc. Res.* **117**, 863–875 (2021).
19. Goodrich, J. K. et al. Conducting a microbiome study. *Cell* **158**, 250–262 (2014).
20. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* **8**, 1561 (2017).
21. Kim, D. et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52 (2017).
22. Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
23. Rahman, G. et al. Determination of effect sizes for power analysis for microbiome studies using large microbiome databases. *Genes* **14**, 1239 (2023).
24. Ferdous, T. et al. The rise to power of the microbiome: power and sample size calculation for microbiome studies. *Mucosal Immunol.* **15**, 1060–1070 (2022).
25. McGuire, M. K. & McGuire, M. A. Got bacteria? The astounding, yet not-so-surprising, microbiome of human milk. *Curr. Opin. Biotechnol.* **44**, 63–68 (2017).
26. Heymann, C. J. F., Bard, J.-M., Heymann, M.-F., Heymann, D. & Bobin-Dubigeon, C. The intratumoral microbiome: characterization methods and functional impact. *Cancer Lett.* **522**, 63–79 (2021).
27. Dickson, R. P., Erb-Downward, J. R., Martinez, F. J. & Huffnagle, G. B. The microbiome and the respiratory tract. *Annu. Rev. Physiol.* **78**, 481–504 (2016).
28. France, M. T. et al. Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. *Genome Biol.* **23**, 66 (2022).
29. Liu, F. et al. Comparative metagenomic analysis of the vaginal microbiome in healthy women. *Synth. Syst. Biotechnol.* **6**, 77–84 (2021).
30. Thomas-White, K., Brady, M., Wolfe, A. J. & Mueller, E. R. The bladder is not sterile: history and current discoveries on the urinary microbiome. *Curr. Bladder Dysfunct. Rep.* **11**, 18–24 (2016).

31. Perez, G. I. P. et al. Body site is a more determinant factor than human population diversity in the healthy skin microbiome. *PLoS ONE* **11**, e0151990 (2016).
32. Marotz, C. A. et al. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42 (2018).
33. Ilett, E. E. et al. Gut microbiome comparability of fresh-frozen versus stabilized-frozen samples from hospitalized patients using 16S rRNA gene and shotgun metagenomic sequencing. *Sci. Rep.* **9**, 13351 (2019).
34. Byrd, D. A. et al. Comparison of methods to collect fecal samples for microbiome studies using whole-genome shotgun metagenomic sequencing. *mSphere* https://doi.org/10.1128/msphere.00827-19 (2020).
35. Andreani, N. A., Donaldson, C. J. & Goddard, M. A reasonable correlation between cloacal and cecal microbiomes in broiler chickens. *Poult. Sci.* **99**, 6062–6070 (2020).
36. Videvall, E., Strandh, M., Engelbrecht, A., Cloete, S. & Cornwallis, C. K. Measuring the gut microbiome in birds: comparison of faecal and cloacal sampling. *Mol. Ecol. Resour.* **18**, 424–434 (2018).
37. Ericsson, A. C. et al. The influence of caging, bedding, and diet on the composition of the microbiota in different regions of the mouse gut. *Sci. Rep.* **8**, 4065 (2018).
38. Jiang, W. et al. Optimized DNA extraction and metagenomic sequencing of airborne microbial communities. *Nat. Protoc.* **10**, 768–779 (2015).
39. Gusareva, E. S. et al. Taxonomic composition and seasonal dynamics of the air microbiome in West Siberia. *Sci. Rep.* **10**, 21515 (2020).
40. Leung, M. H. Y. et al. Characterization of the public transit air microbiome and resistome reveals geographical specificity. *Microbiome* **9**, 112 (2021).
41. Behzad, H., Gojobori, T. & Mineta, K. Challenges and opportunities of airborne metagenomics. *Genome Biol. Evol.* **7**, 1216–1226 (2015).
42. James, G. L. et al. Metagenomic datasets of air samples collected during episodes of severe smoke-haze in Malaysia. *Data Brief.* **36**, 107124 (2021).
43. Tong, X. et al. Metagenomic insights into the microbial communities of inert and oligotrophic outdoor pier surfaces of a coastal city. *Microbiome* **9**, 213 (2021).
44. Acharya, K. et al. Metagenomic water quality monitoring with a portable laboratory. *Water Res.* **184**, 116112 (2020).
45. Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* **4**, 1–15 (2021).
46. Poulsen, C. S., Kaas, R. S., Aarestrup, F. M. & Pamp, S. J. Standard sample storage conditions have an impact on inferred microbiome composition and antimicrobial resistance patterns. *Microbiol. Spectr.* **9**, e0138721 (2021).
47. Hickl, O. et al. Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut microbiome. *Microorganisms* **7**, 367 (2019).
48. Sinha, R. et al. Collecting fecal samples for microbiome analyses in epidemiology studies. *Cancer Epidemiol. Biomarkers Prev.* **25**, 407–416 (2016).
49. Byrd, D. A. et al. Reproducibility, stability, and accuracy of microbial profiles by fecal sample collection method in three distinct populations. *PLoS ONE* **14**, e0224757 (2019).
50. Michael, A. G., Zoe, A. P. & Christina, A. K. Comparison of DNA preservation methods for environmental bacterial community samples. *FEMS Microbiol. Ecol.* **83**, 468–477 (2013).
51. Yang, F. et al. Assessment of fecal DNA extraction protocols for metagenomic studies. *GigaScience* **9**, giaa071 (2020).
52. Yang, L. et al. Preservation of the fecal samples at ambient temperature for microbiota analysis with a cost-effective and reliable stabilizer EffcGut. *Sci. Total. Environ.* **741**, 140423 (2020).
53. Chen, C.-C. et al. Comparison of DNA stabilizers and storage conditions on preserving fecal microbiota profiles. *J. Formos. Med. Assoc.* **119**, 1791–1798 (2020).
54. Fabre, A.-L., Colotte, M., Luis, A., Tuffet, S. & Bonnet, J. An efficient method for long-term room temperature storage of RNA. *Eur. J. Hum. Genet.* **22**, 379–385 (2014).
55. Choudhury, R., Middelkoop, A., Bolhuis, J. E. & Kleerebezem, M. Legitimate and reliable determination of the age-related intestinal microbiome in young piglets; rectal swabs and fecal samples provide comparable insights. *Front. Microbiol.* **10**, 1886 (2019).
56. Hendriksen, R. S. et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).
57. Teytelman, L., Stoliartchouk, A., Kindler, L. & Hurwitz, B. L. protocols.io: virtual communities for protocol development and discussion. *PLoS Biol.* **14**, e1002538 (2016).
58. Marcus, E. A. STAR is born. *Cell* **166**, 1059–1060 (2016).
59. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
60. Greathouse, K. L., Sinha, R. & Vogtmann, E. DNA extraction for human microbiome studies: the issue of standardization. *Genome Biol.* **20**, 212 (2019).
61. Sui, H. et al. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. *Front. Microbiol.* **11**, 953 (2020).
62. Peng, Z. et al. Comparative analysis of sample extraction and library construction for shotgun metagenomics. *Bioinform. Biol. Insights* **14**, 1177932220915459 (2020).
63. Rehner, J. et al. Systematic cross-biospecimen evaluation of DNA extraction kits for long- and short-read multi-metagenomic sequencing studies. *Genomics Proteom. Bioinform.* **20**, 405–417 (2022).
64. Shaffer, J. P. et al. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat. Microbiol.* **7**, 2128–2150 (2022).
65. Tourlousse, D. M. et al. Validation and standardization of DNA extraction and library construction methods for metagenomics-based human fecal microbiome measurements. *Microbiome* **9**, 95 (2021).
66. d'Humières, C. et al. The potential role of clinical metagenomics in infectious diseases: therapeutic perspectives. *Drugs* **81**, 1453–1466 (2021).
67. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).
68. Head, S. R. et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* **56**, 61–77 (2014).
69. Modi, A., Vai, S., Caramelli, D. & Lari, M. in *Bacterial Pangenomics: Methods and Protocols* (eds Mengoni, A., Bacci, G. & Fondi, M.) 15–42 (Springer, 2021).
70. Barba, M., Czosnek, H. & Hadidi, A. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* **6**, 106–136 (2014).
71. 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. Genomics* **8**, 000794 (2022).
72. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
73. Gounot, J.-S. et al. Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nat. Commun.* **13**, 6044 (2022).
74. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinform.* **3**, lqab019 (2021).
75. Sanderson, N. D. et al. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microb. Genomics* **10**, 001246 (2024).
76. Bogaerts, B. et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *J. Clin. Microbiol.* **62**, e0157623 (2024).
77. Eisenhofer, R. et al. A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. *Microbiol. Spectr.* **12**, e0359023 (2024).
78. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1136 (2019).
79. Johnson, M. S., Venkataram, S. & Kryazhimskiy, S. Best practices in designing, sequencing, and identifying random DNA barcodes. *J. Mol. Evol.* **91**, 263–280 (2023).
80. Werner, J. J., Zhou, D., Caporaso, J. G., Knight, R. & Angenent, L. T. Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J.* **6**, 1273–1276 (2012).
81. Karamitros, T. & Magiorkinis, G. in *Next Generation Sequencing: Methods and Protocols* (eds Head, S. R., Ordoukhanian, P. & Salomon, D. R.) 43–51 (Springer, 2018).
82. Kong, N. et al. Automation of PacBio SMRTbell NGS library preparation for bacterial genome sequencing. *Stand. Genomic Sci.* **12**, 27 (2017).
83. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
84. Nishii, K. et al. A high quality, high molecular weight DNA extraction method for PacBio HiFi genome sequencing of recalcitrant plants. *Plant Methods* **19**, 41 (2023).
85. Kim, C. Y., Ma, J. & Lee, I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat. Commun.* **13**, 6367 (2022).
86. Du, Y. & Sun, F. MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. *Nat. Commun.* **14**, 6231 (2023).
87. Beitel, C. W. et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
88. Li, W. & Godzik, A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
89. Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinform.* **12**, 271 (2011).
90. Tavakolian, N., Frazão, J. G., Bendixsen, D., Stelkens, R. & Li, C.-B. Shepherd: accurate clustering for correcting DNA barcode errors. *Bioinformatics* **38**, 3710–3716 (2022).
91. Roehr, J. T., Dieterich, C. & Reinert, K. Flexbar 3.0—SIMD and multicore parallelization. *Bioinformatics* **33**, 2941–2942 (2017).
92. Busch, A., Brüggemann, M., Ebersberger, S. & Zarnack, K. iCLIP data analysis: a complete pipeline from sequencing reads to RBP binding sites. *Methods* **178**, 49–62 (2020).
93. Wilkins, O. G., Capitanchik, C., Luscombe, N. M. & Ule, J. Ultraplex: a rapid, flexible, all-in-one FASTQ demultiplexer. *Wellcome Open Res.* **6**, 141 (2021).
94. König, J. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
95. Kong, Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* **98**, 152–153 (2011).
96. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
97. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
98. Murray, K. D. & Borevitz, J. O. Axe: rapid, competitive sequence read demultiplexing using a trie. *Bioinformatics* **34**, 3924–3925 (2018).
99. Liu, D. Fuzzysplit: demultiplexing and trimming sequenced DNA with a declarative language. *PeerJ* **7**, e7170 (2019).
100. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
101. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
102. 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).

103. Fukasawa, Y., Ermini, L., Wang, H., Carty, K. & Cheung, M.-S. LongQC: a quality control tool for third generation sequencing long read data. *Genes Genomes Genet.* **10**, 1193–1196 (2020).
104. 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).
105. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).
106. Krehenwinkel, H. et al. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* **8**, giz006 (2019).
107. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
108. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
109. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
110. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat. Methods* **19**, 671–674 (2022).
111. Benoit, G. et al. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-023-01983-6 (2024).
112. Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M. & Yorke, J. A. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**, 3363–3369 (2004).
113. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
114. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
115. 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).
116. 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).
117. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
118. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
119. Brown, C. L. et al. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Sci. Rep.* **11**, 3753 (2021).
120. Chen, Z., Erickson, D. L. & Meng, J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics* **21**, 631 (2020).
121. Liu, L. et al. Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. *Microbiome* **9**, 205 (2021).
122. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
123. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 101 (2021).
124. Hyeon, J.-Y., Mann, D. A., Townsend, A. M. & Deng, X. Quasi-metagenomic analysis of *Salmonella* from food and environmental samples. *J. Vis. Exp.* https://doi.org/10.3791/58612 (2018).
125. Cilibrasi, R., van Iersel, L., Kelk, S. & Tromp, J. in *Algorithms in Bioinformatics* (eds Casadio, R. & Myers, G.) 128–139 (Springer, 2005).
126. Nicholls, S. M. et al. On the complexity of haplotyping a microbial community. *Bioinformatics* **37**, 1360–1366 (2021).
127. 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 Bioinform.* **23**, 541 (2022).
128. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
129. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
130. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
131. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
132. 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).
133. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
134. 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).
135. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12**, S4 (2011).
136. Darling, A. E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
137. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-023-01688-w (2023).
138. Martínez-Porchas, M., Villalpando-Canchola, E. & Vargas-Albores, F. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon* **2**, e00170 (2016).
139. Alser, M. et al. Technology dictates algorithms: recent developments in read alignment. *Genome Biol.* **22**, 249 (2021).
140. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
141. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
142. Ondov, B. D. et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232 (2019).
143. Koslicki, D., White, S., Ma, C. & Novikov, A. YACHT: an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample. *Bioinformatics* **40**, btae047 (2024).
144. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
145. LaPierre, N., Alser, M., Eskin, E., Koslicki, D. & Mangul, S. Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome Biol.* **21**, 242 (2020).
146. Liu, S. & Koslicki, D. CMash: fast, multi-resolution estimation of *k*-mer-based Jaccard and containment indices. *Bioinformatics* **38**, i28–i35 (2022).
147. Agustinho, D. P. et al. Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat. Methods* **21**, 954–966 (2024).
148. Chen, L. et al. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).
149. Tedersoo, L., Albertsen, M., Anslan, S. & Callahan, B. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl. Environ. Microbiol.* **87**, e0062621 (2021).
150. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).
151. Boolchandani, M., Patel, S. & Dantas, G. in *Antibiotics: Methods and Protocols* (ed. Sass, P.) 307–329 (Springer, 2017).
152. Healy, F. G. et al. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl. Microbiol. Biotechnol.* **43**, 667–674 (1995).
153. Williamson, L. L. et al. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl. Environ. Microbiol.* **71**, 6335–6344 (2005).
154. Riesenfeld, C. S., Goodman, R. M. & Handelsman, J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ. Microbiol.* **6**, 981–989 (2004).
155. Turnbaugh, P. J. et al. The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
156. Proctor, L. M. et al. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
157. Dabdoub, S. M., Ganesan, S. M. & Kumar, P. S. Comparative metagenomics reveals taxonomically idiosyncratic yet functionally congruent communities in periodontitis. *Sci. Rep.* **6**, 38993 (2016).
158. Badger, J. H. & Olsen, G. J. CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512–524 (1999).
159. Frishman, D., Mironov, A., Mewes, H.-W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26**, 2941–2947 (1998).
160. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).
161. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
162. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* **40**, e9 (2012).
163. Borodovsky, M. & McIninch, J. GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123–133 (1993).
164. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
165. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, W451–W454 (2005).
166. Krause, L. et al. GISMO—gene identification using a support vector machine for ORF classification. *Nucleic Acids Res.* **35**, 540–549 (2007).
167. DeCaprio, D. et al. Conrad: gene prediction using conditional random fields. *Genome Res.* **17**, 1389–1398 (2007).
168. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* **15**, 387–396 (2008).
169. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
170. Al-Ajlan, A. & El Allali, A. CNN-MGP: convolutional neural networks for metagenomics gene prediction. *Interdiscip. Sci. Comput. Life Sci* **11**, 628–635 (2019).

171. Zhang, S.-W., Jin, X.-Y. & Zhang, T. Gene prediction in metagenomic fragments with deep learning. *BioMed. Res. Int.* **2017**, e4740354 (2017).
172. Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A. & Creevey, C. J. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* **38**, 1198–1207 (2022).
173. Dong, X. & Strous, M. An integrated pipeline for annotation and visualization of metagenomic contigs. *Front. Genet.* **10**, 999 (2019).
174. Franzosa, E. A. et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13**, 360–372 (2015).
175. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
176. Maranga, M. et al. Comprehensive functional annotation of metagenomes and microbial genomes using a deep learning-based method. *mSystems* **8**, e0117822 (2023).
177. Yue, Y. et al. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinform.* **21**, 334 (2020).
178. Niu, S.-Y. et al. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief. Bioinform.* **19**, 1415–1429 (2018).
179. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
180. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
181. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
182. Galperin, M. Y. et al. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
183. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkae909 (2024).
184. Anand, S., Kuntal, B. K., Mohapatra, A., Bhatt, V. & Mande, S. S. FunGeCo: a web-based tool for estimation of functional potential of bacterial genomes and microbiomes using gene context information. *Bioinformatics* **36**, 2575–2577 (2020).
185. Ruperti, F. et al. Cross-phyla protein annotation by structural prediction and alignment. *Genome Biol.* **24**, 113 (2023).
186. Scheibenreif, L., Littmann, M., Orengo, C. & Rost, B. FunFam protein families improve residue level molecular function prediction. *BMC Bioinform.* **20**, 400 (2019).
187. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
188. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
189. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
190. Vanni, C. et al. Unifying the known and unknown microbial coding sequence space. *eLife* **11**, e67667 (2022).
191. Meyer, F. et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **9**, 386 (2008).
192. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
193. Hou, Q. et al. Using metagenomic data to boost protein structure prediction and discovery. *Comput. Struct. Biotechnol. J.* **20**, 434–442 (2022).
194. Wang, Q., Han, Y., Lan, S. & Hu, C. Metagenomic insight into patterns and mechanism of nitrogen cycle during biocrust succession. *Front. Microbiol.* **12**, 633428 (2021).
195. Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems* **4**, e00332-18 (2019).
196. Haiminen, N., Utro, F., Seabolt, E. & Parida, L. Functional profiling of COVID-19 respiratory tract microbiomes. *Sci. Rep.* **11**, 6433 (2021).
197. Keegan, K. P., Glass, E. M. & Meyer, F. in *Microbial Environmental Genomics (MEG)* (eds Martin, F. & Uroz, S.) 207–233 (Springer, 2016).
198. Wilke, A. et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinform.* **13**, 141 (2012).
199. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
200. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
201. Aramaki, T. et al. KOfamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
202. Coudert, E. et al. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**, btac793 (2023).
203. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
204. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
205. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
206. Pascal Andreu, V. et al. gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01675-1 (2023).
207. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
208. Arkin, A. P. et al. KBase: the United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).
209. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
210. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
211. DeGruttola, A. K., Low, D., Mizoguchi, A. & Mizoguchi, E. Current understanding of dysbiosis in disease in human and animal models. *Inflamm. Bowel Dis.* **22**, 1137–1150 (2016).
212. Liu, L. et al. Gut microbiota and its metabolites in depression: from pathogenesis to treatment. *eBioMedicine* **90**, 104527 (2023).
213. Zhang, X. et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905 (2015).
214. Kitsios, G. D. et al. Longitudinal multicompartment characterization of host–microbiota interactions in patients with acute respiratory failure. *Nat. Commun.* **15**, 4708 (2024).
215. Cheng, M. et al. Deep longitudinal lower respiratory tract microbiome profiling reveals genome-resolved functional and evolutionary dynamics in critical illness. *Nat. Commun.* **15**, 8361 (2024).
216. Yin, L. et al. Association between gut bacterial diversity and mortality in septic shock patients: a cohort study. *Med. Sci. Monit.* **25**, 7376–7382 (2019).
217. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
218. Ai, D. et al. Integrated metagenomic data analysis demonstrates that a loss of diversity in oral microbiota is associated with periodontitis. *BMC Genomics* **18**, 1041 (2017).
219. Allin, K. H. et al. Aberrant intestinal microbiota in individuals with prediabetes. *Diabetologia* **61**, 810–820 (2018).
220. Dey, N., Soergel, D. A., Repo, S. & Brenner, S. E. Association of gut microbiota with post-operative clinical course in Crohn's disease. *BMC Gastroenterol.* **13**, 131 (2013).
221. Plassais, J. et al. Gut microbiome alpha-diversity is not a marker of Parkinson's disease and multiple sclerosis. *Brain Commun.* **3**, fcab113 (2021).
222. Simpson, C. A. et al. The gut microbiota in anxiety and depression—a systematic review. *Clin. Psychol. Rev.* **83**, 101943 (2021).
223. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
224. Zielińska, K. et al. Healthy microbiome—moving towards functional interpretation. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.04.569909 (2024).
225. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
226. Wu, J. et al. Metagenomic next-generation sequencing in detecting pathogens in pediatric oncology patients with suspected bloodstream infections. *Pediatr. Res.* **95**, 843–851 (2024).
227. Tan, C. C. S., Acman, M., van Dorp, L. & Balloux, F. Metagenomic evidence for a polymicrobial signature of sepsis. *Microb. Genom.* **7**, 000642 (2021).
228. Stacy, A., McNally, L., Darch, S. E., Brown, S. P. & Whiteley, M. The biogeography of polymicrobial infection. *Nat. Rev. Microbiol.* **14**, 93–105 (2016).
229. Stacy, A. et al. Bacterial fight-and-flight responses enhance virulence in a polymicrobial infection. *Proc. Natl Acad. Sci. USA* **111**, 7819–7824 (2014).
230. Peters, B. M., Jabra-Rizk, M. A., O'May, G. A., Costerton, J. W. & Shirtliff, M. E. Polymicrobial interactions: impact on pathogenesis and human disease. *Clin. Microbiol. Rev.* **25**, 193–213 (2012).
231. Liu, Z. et al. Metagenomic and metatranscriptomic analyses reveal activity and hosts of antibiotic resistance genes in activated sludge. *Environ. Int.* **129**, 208–220 (2019).
232. Sukhum, K. V., Diorio-Toth, L. & Dantas, G. Genomic and metagenomic approaches for predictive surveillance of emerging pathogens and antibiotic resistance. *Clin. Pharmacol. Ther.* **106**, 512–524 (2019).
233. Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
234. Dapa, T. et al. Within-host evolution of the gut microbiome. *Curr. Opin. Microbiol.* **71**, 102258 (2023).
235. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
236. de Nies, L. et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 49 (2021).
237. Levade, I. et al. Predicting *Vibrio cholerae* infection and disease severity using metagenomics in a prospective cohort study. *J. Infect. Dis.* **223**, 342–351 (2021).
238. Ianiro, G. et al. Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. *Nat. Med.* **28**, 1913–1923 (2022).
239. van Schaik, W. The human gut resistome. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140087 (2015).
240. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
241. Wilson, M. R. et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
242. Ibberson, C. B., Barraza, J. P., Holmes, A. L., Cao, P. & Whiteley, M. Precise spatial structure impacts antimicrobial susceptibility of *S. aureus* in polymicrobial wound infections. *Proc. Natl Acad. Sci. USA* **119**, e2212340119 (2022).

243. Ramsey, M. M. & Whiteley, M. Polymicrobial interactions stimulate resistance to host innate immunity through metabolite perception. *Proc. Natl Acad. Sci. USA* **106**, 1578–1583 (2009).
244. Rogers, G. B. et al. Revealing the dynamics of polymicrobial infections: implications for antibiotic therapy. *Trends Microbiol.* **18**, 357–364 (2010).
245. Smith, A. B. et al. Enterococci enhance *Clostridioides difficile* pathogenesis. *Nature* **611**, 780–786 (2022).
246. Qin, C. et al. Diagnostic value of metagenomic next-generation sequencing in sepsis and bloodstream infection. *Front. Cell. Infect. Microbiol.* **13**, 1117987 (2023).
247. Schlaberg, R. et al. Viral pathogen detection by metagenomics and pan-viral group polymerase chain reaction in children with pneumonia lacking identifiable etiology. *J. Infect. Dis.* **215**, 1407–1415 (2017).
248. Zhou, Y. et al. Metagenomic approach for identification of the pathogens associated with diarrhea in stool specimens. *J. Clin. Microbiol.* **54**, 368–375 (2016).
249. Miller, S. et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* **29**, 831–842 (2019).
250. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
251. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
252. Quinn, R. A. et al. Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *mBio* https://doi.org/10.1128/mbio.00956-13 (2014).
253. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
254. Peddu, V. et al. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. *Clin. Chem.* **66**, 966–972 (2020).
255. Carbo, E. C. et al. Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness. *J. Clin. Virol.* **131**, 104594 (2020).
256. Castañeda-Mogollón, D. et al. A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity. *J. Clin. Virol.* **145**, 105025 (2021).
257. Ko, K. K. K., Chng, K. R. & Nagarajan, N. Metagenomics-enabled microbial surveillance. *Nat. Microbiol.* **7**, 486–496 (2022).
258. Laudadio, I., Fulci, V., Stronati, L. & Carissimi, C. Next-generation metagenomics: methodological challenges and opportunities. *OMICS J. Integr. Biol.* **23**, 327–333 (2019).
259. Brito, I. L. & Alm, E. J. Tracking strains in the microbiome: insights from metagenomics and models. *Front. Microbiol.* **7**, 712 (2016).
260. Nutman, A. & Marchaim, D. How to: molecular investigation of a hospital outbreak. *Clin. Microbiol. Infect.* **25**, 688–695 (2019).
261. Zhao, C., Dimitrov, B., Goldman, M., Nayfach, S. & Pollard, K. S. MIDAS2: metagenomic intra-species diversity analysis system. *Bioinformatics* **39**, btac713 (2023).
262. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
263. Podlesny, D. et al. Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. *Microbiome* **10**, 53 (2022).
264. Chng, K. R. et al. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat. Med.* **26**, 941–951 (2020).
265. Lax, S. et al. Bacterial colonization and succession in a newly opened hospital. *Sci. Transl. Med.* **9**, eaah6500 (2017).
266. Maruyama, R. et al. Metagenomic analysis of the microbial communities and associated network of nitrogen metabolism genes in the Ryukyu limestone aquifer. *Sci. Rep.* **14**, 4356 (2024).
267. Tran, P. Q. et al. Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J.* **15**, 1971–1986 (2021).
268. Mani, I. in *Bioremediation of Pollutants* (eds Pandey, V. C. & Singh, V.) 275–285 (Elsevier, 2020).
269. Raynaud, X. & Nunan, N. Spatial ecology of bacteria at the microscale in soil. *PLoS ONE* **9**, e87217 (2014).
270. Richardson, A. E. & Simpson, R. J. Soil microorganisms mediating phosphorus availability update on microbial phosphorus. *Plant. Physiol.* **156**, 989–996 (2011).
271. Medić, A. B. & Karadžić, I. M. *Pseudomonas* in environmental bioremediation of hydrocarbons and phenolic compounds—key catabolic degradation enzymes and new analytical platforms for comprehensive investigation. *World J. Microbiol. Biotechnol.* **38**, 165 (2022).
272. Gupta, V. K. & Rastogi, A. Biosorption of lead from aqueous solutions by green algae *Spirogyra* species: kinetics and equilibrium studies. *J. Hazard. Mater.* **152**, 407–414 (2008).
273. Olguín, E. J. Phycoremediation: key issues for cost-effective nutrient removal processes. *Biotechnol. Adv.* **22**, 81–91 (2003).
274. Lovley, D. R. Cleaning up with genomics: applying molecular biology to bioremediation. *Nat. Rev. Microbiol.* **1**, 35–44 (2003).
275. Kachienga, L., Jitendra, K. & Momba, M. Metagenomic profiling for assessing microbial diversity and microbial adaptation to degradation of hydrocarbons in two South African petroleum-contaminated water aquifers. *Sci. Rep.* **8**, 7564 (2018).
276. Akoijam, N. & Joshi, S. R. in *Molecular Genetics and Genomics Tools in Biodiversity Conservation* (eds Kumar, A., Choudhury, B., Dayanandan, S. & Khan, M. L.) 31–61 (Springer Nature, 2022).
277. Wei, F. et al. Conservation metagenomics: a new branch of conservation biology. *Sci. China Life Sci* **62**, 168–178 (2019).
278. Zheng, D. et al. Metagenomics highlights the impact of climate and human activities on antibiotic resistance genes in China's estuaries. *Environ. Pollut.* **301**, 119015 (2022).
279. Li, J. et al. Microbiome engineering: a promising approach to improve coral health. *Engineering* **28**, 105–116 (2023).
280. Wrighton, K. H. Discovering antibiotics through soil metagenomics. *Nat. Rev. Drug. Discov.* **17**, 241–241 (2018).
281. Miethke, M. et al. Towards the sustainable discovery and development of new antibiotics. *Nat. Rev. Chem.* **5**, 726–749 (2021).
282. Shin, S.-K. et al. Metagenomic insights into the bioaerosols in the indoor and outdoor environments of childcare facilities. *PLoS ONE* **10**, e0126960 (2015).
283. Singh, B. K., Trivedi, P., Egidi, E., Macdonald, C. A. & Delgado-Baquerizo, M. Crop microbiome and sustainable agriculture. *Nat. Rev. Microbiol.* **18**, 601–602 (2020).
284. Berihu, M. et al. A framework for the targeted recruitment of crop-beneficial soil taxa based on network analysis of metagenomics data. *Microbiome* **11**, 8 (2023).
285. Vimal, S. R., Singh, J. S., Kumar, A. & Prasad, S. M. Plant genotype-microbiome engineering as nature-based solution (NbS) for regeneration of stressed agriculture: a review. *Sci. Hortic.* **321**, 112258 (2023).
286. Kumar, S., Diksha, Sindhu, S. S. & Kumar, R. Biofertilizers: an ecofriendly technology for nutrient recycling and environmental sustainability. *Curr. Res. Microb. Sci.* **3**, 100094 (2022).
287. Li, R. et al. Metagenomic analysis exploring taxonomic and functional diversity of soil microbial communities in sugarcane fields applied with organic fertilizer. *BioMed. Res. Int.* **2020**, e9381506 (2020).
288. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
289. Scholz, M. B., Lo, C.-C. & Chain, P. S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* **23**, 9–15 (2012).
290. ten Hoopen, P. et al. The metagenomic data life-cycle: standards and best practices. *Gigascience* **6**, 1–11 (2017).
291. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
292. Field, D. et al. The Genomic Standards Consortium. *PLoS Biol.* **9**, e1001088 (2011).
293. Barrett, T. et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40**, D57–D63 (2012).
294. Haft, D. H. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
295. Birney, E. et al. An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004).
296. Gillespie, J. J. et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**, 4286–4298 (2011).
297. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
298. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
299. Basenko, E. Y. et al. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. *J. Fungi* **4**, 39 (2018).
300. Grigoriev, I. V. et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
301. Ahrendt, S. R., Mondo, S. J., Haridas, S. & Grigoriev, I. V. in *Microbial Environmental Genomics (MEG)* (eds Martin, F. & Uroz, S.) 271–291 (Springer, 2023).
302. Olson, R. D. et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkac1003 (2022).
303. Uchiyama, I. MBGD: Microbial Genome Database for comparative analysis. *Nucleic Acids Res.* **31**, 58–62 (2003).
304. Kim, C. Y. et al. Human Reference Gut Microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med.* **13**, 134 (2021).
305. Yooseph, S. et al. A metagenomic framework for the study of airborne microbial communities. *PLoS ONE* **8**, e81862 (2013).
306. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
307. Data sharing and the future of science. *Nat. Commun.* **9**, 2817 (2018).
308. Denk, F. Don't let useful data go to waste. *Nature* **543**, 7 (2017).
309. Torres, P. J., Edwards, R. A. & McNair, K. A. PARTIE: a partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. *Bioinformatics* **33**, 2389–2391 (2017).
310. Vuong, P., Wise, M. J., Whiteley, A. S. & Kaur, P. Ten simple rules for investigating (meta) genomic data from environmental ecosystems. *PLoS Comput. Biol.* **18**, e1010675 (2022).
311. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
312. Boon, E. et al. Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol. Rev.* **38**, 90–118 (2014).

313. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
314. Hover, B. M. et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
315. Li, X. & Qin, L. Metagenomics-based drug discovery and marine microbial diversity. *Trends Biotechnol.* **23**, 539–543 (2005).
316. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
317. Swetnam, T. L. et al. CyVerse: cyberinfrastructure for open science. *PLoS Comput. Biol.* **20**, e1011270 (2024).
318. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
319. Araújo-Pérez, F. et al. Differences in microbial signatures between rectal mucosal biopsies and rectal swabs. *Gut Microbes* **3**, 530–535 (2012).
320. Shaw, A. G. et al. Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room? *Microbiome* **4**, 40 (2016).
321. Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5**, 16350 (2015).
322. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
323. Tilg, H. & Moschen, A. R. Microbiota and diabetes: an evolving relationship. *Gut* **63**, 1513–1521 (2014).
324. Vujkovic-Cvijin, I. et al. Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
325. Zhernakova, D. V. et al. Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813–821 (2024).
326. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat. Genet.* **54**, 134–142 (2022).
327. Liu, X. et al. Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. *Cell Discov.* **7**, 1–16 (2021).
328. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 24 (2016).
329. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
330. 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).
331. Knights, D. et al. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).
332. Ganda, E. et al. DNA extraction and host depletion methods significantly impact and potentially bias bacterial detection in a biological fluid. *mSystems* https://doi.org/10.1128/msystems.00619-21 (2021).
333. Yap, M. et al. Evaluation of methods for the reduction of contaminating host reads when performing shotgun metagenomic sequencing of the milk microbiome. *Sci. Rep.* **10**, 21665 (2020).
334. Bruggeling, C. E. et al. Optimized bacterial DNA isolation method for microbiome analysis of human tissues. *MicrobiologyOpen* **10**, e1191 (2021).
335. Eisenhofer, R. et al. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* **27**, 105–117 (2019).
336. Nelson, M. T. et al. Human and extracellular DNA depletion for metagenomic analysis of complex clinical infection samples yields optimized viable microbiome profiles. *Cell Rep.* **26**, 2227–2240.e5 (2019).
337. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* **9**, e97876 (2014).
338. Carini, P. et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* **2**, 1–6 (2016).
339. Lutz, K. A., Wang, W., Zdepski, A. & Michael, T. P. Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnol.* **11**, 54 (2011).
340. Koutsovoulos, G. et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc. Natl Acad. Sci. USA* **113**, 5053–5058 (2016).
341. Thoendel, M. et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J. Microbiol. Methods* **127**, 141–145 (2016).
342. Pearce, M. M. et al. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* https://doi.org/10.1128/mbio.01283-14 (2014).
343. Natalini, J. G., Singh, S. & Segal, L. N. The dynamic lung microbiome in health and disease. *Nat. Rev. Microbiol.* **21**, 222–235 (2023).
344. Luhung, I. et al. Experimental parameters defining ultra-low biomass bioaerosol analysis. *npj Biofilms Microbiomes* **7**, 1–11 (2021).
345. Hasrat, R. et al. Benchmarking laboratory processes to characterise low-biomass respiratory microbiota. *Sci. Rep.* **11**, 17148 (2021).
346. Simpson, A. C. et al. Analysis of microbiomes from ultra-low biomass surfaces using novel surface sampling and nanopore sequencing. *J. Biomol. Tech.* **34**, 3fc1f5fe.bac4a5b3 (2023).
347. Carey-Ann, B. D. & Carroll, K. C. Diagnosis of *Clostridium difficile* infection: an ongoing conundrum for clinicians and for clinical laboratories. *Clin. Microbiol. Rev.* **26**, 604–630 (2013).
348. Aagaard, K. et al. The placenta harbors a unique microbiome. *Sci. Transl. Med.* **6**, 237ra65 (2014).
349. Antony, K. M. et al. The preterm placental microbiome varies in association with excess maternal gestational weight gain. *Am. J. Obstet. Gynecol.* **212**, 653.e1–e16 (2015).
350. Amarasekara, R., Jayasekara, R. W., Senanayake, H. & Dissanayake, V. H. W. Microbiome of the placenta in pre-eclampsia supports the role of bacteria in the multifactorial cause of pre-eclampsia. *J. Obstet. Gynaecol. Res.* **41**, 662–669 (2015).
351. O'Callaghan, J. et al. Re-assessing microbiomes in the low-biomass reproductive niche. *BJOG* **127**, 147–158 (2020).
352. Kennedy, K. M et al. Fetal meconium does not have a detectable microbiota before birth. *Nat. Microbiol.* **6**, 865–873 (2021).
353. de Goffau, M. C. et al. Human placenta has no microbiome but can contain potential pathogens. *Nature* **572**, 329–334 (2019).
354. Gschwind, R. et al. Evidence for contamination as the origin for bacteria found in human placenta rather than a microbiota. *PLoS ONE* **15**, e0237232 (2020).
355. Lauder, A. P. et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 29 (2016).
356. Blaser, M. J. et al. Lessons learned from the prenatal microbiome controversy. *Microbiome* **9**, 8 (2021).
357. 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).
358. Nagarajan, N. & Pop, M. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J. Comput. Biol.* **16**, 897–908 (2009).
359. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
360. Wu, R. et al. Hi-C metagenome sequencing reveals soil phage–host interactions. *Nat. Commun.* **14**, 7666 (2023).
361. Šimková, H., Câmara, A. S. & Mascher, M. Hi-C techniques: from genome assemblies to transcription regulation. *J. Exp. Botany* **75**, 5357–5365 (2024).
362. Sonets, I. V. et al. Hi-C metagenomics facilitate comparative genome analysis of bacteria and yeast from spontaneous beer and cider. *Food Microbiol.* **121**, 104520 (2024).
363. Du, Y. & Sun, F. HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome Biol.* **23**, 63 (2022).
364. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
365. Parker, C. T., Tindall, B. J. & Garrity, G. M. (eds) International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **69**, S1–S111 (2019).
366. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
367. Lloyd, K. G. & Tahon, G. Science depends on nomenclature, but nomenclature is not science. *Nat. Rev. Microbiol.* **20**, 123–124 (2022).
368. Jacobson, D. K. et al. Analysis of global human gut metagenomes shows that metabolic resilience potential for short-chain fatty acid production is strongly influenced by lifestyle. *Sci. Rep.* **11**, 1724 (2021).
369. Rusch, D. B et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
370. Vogel, T. M. et al. TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.* **7**, 252 (2009).
371. Ryon, K. A. et al. A history of the MetaSUB consortium: tracking urban microbes around the globe. *iScience* **25**, 104993 (2022).
372. Methé, B. A. et al. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
373. Engel, P. et al. The bee microbiome: impact on bee health and model for evolution and ecology of host–microbe interactions. *mBio* **7**, e02164-15 (2016).
374. Mangul, S. et al. Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 1393 (2019).
375. Meyer, F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
376. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
377. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
378. Alvarez-Jarreta, J. et al. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023. *Nucleic Acids Res.* **52**, D808–D816 (2024).
379. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
380. Uchiyama, I., Mihara, M., Nishide, H., Chiba, H. & Kato, M. MBGD update 2018: Microbial Genome Database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res.* **47**, D382–D389 (2019).
381. Maillet, A. et al. Impact of DNA extraction and sampling methods on bacterial communities monitored by 16S rDNA metabarcoding in cold-smoked salmon and processing plant surfaces. *Food Microbiol.* **95**, 103705 (2021).

382. Liang, C. et al. Effects of different storage temperatures on the physicochemical properties and bacterial community structure of fresh lamb meat. *Food Sci. Anim. Resour.* **41**, 509 (2021).
383. Mansur, A. R. et al. Comparative evaluation of spoilage-related bacterial diversity and metabolite profiles in chilled beef stored under air and vacuum packaging. *Food Microbiol.* **77**, 166–172 (2019).
384. Fernandez-Cassi, X. et al. A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water. *Int. J. Food Microbiol.* **257**, 80–90 (2017).
385. Henriet, O., Fourmentin, J., Delincé, B. & Mahillon, J. Exploring the diversity of extremely halophilic archaea in food-grade salts. *Int. J. Food Microbiol.* **191**, 36–44 (2014).
386. Billington, C., Kingsbury, J. M. & Rivas, L. Metagenomics approaches for improving food safety: a review. *J. Food Prot.* **85**, 448–464 (2022).
387. Malla, M. A., Dubey, A., Kumar, A. & Yadav, S. Metagenomic analysis displays the potential predictive biodegradation pathways of the persistent pesticides in agricultural soil with a long record of pesticide usage. *Microbiol. Res.* **261**, 127081 (2022).
388. Jagadesh, M. et al. Revealing the hidden world of soil microbes: metagenomic insights into plant, bacteria, and fungi interactions for sustainable agriculture and ecosystem restoration. *Microbiol. Res.* **285**, 127764 (2024).
389. Medina, J. E. et al. Exploring viral diversity and metagenomics in livestock: insights into disease emergence and spillover risks in cattle. *Vet. Res. Commun.* **48**, 2029–2049 (2024).
390. Buck, M. et al. Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Sci. Data* **8**, 131 (2021).
391. Krinos, A. I. et al. Time-series metagenomics reveals changing protistan ecology of a temperate dimictic lake. *Microbiome* **12**, 133 (2024).
392. Rodríguez-Gijón, A., Hampel, J. J., Dharamshi, J., Bertilsson, S. & Garcia, S. L. Shotgun metagenomes from productive lakes in an urban region of Sweden. *Sci. Data* **10**, 810 (2023).
393. Ijoma, G. N., Ogola, H. J. O., Pillay, P., Tshisekedi, K. A. & Tekere, M. Metagenomics datasets of water and sediments from eutrophication-impacted artificial lakes in South Africa. *Sci. Data* **11**, 456 (2024).
394. Peña-Ocaña, B. A. et al. Metagenomic and metabolic analyses of poly-extreme microbiome from an active crater volcano lake. *Environ. Res.* **203**, 111862 (2022).
395. Mittal, P., Prasoodanan PK, V., Dhakan, D. B., Kumar, S. & Sharma, V. K. Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes. *Environ. Microbiome* **14**, 5 (2019).
396. Zhang, Z.-F., Liu, L.-R., Pan, Y.-P., Pan, J. & Li, M. Long-read assembled metagenomic approaches improve our understanding on metabolic potentials of microbial community in mangrove sediments. *Microbiome* **11**, 188 (2023).
397. Zaouri, N., Jumat, M. R., Cheema, T. & Hong, P.-Y. Metagenomics-based evaluation of groundwater microbial profiles in response to treated wastewater discharge. *Environ. Res.* **180**, 108835 (2020).
398. Brumfield, K. D. et al. A comparative analysis of drinking water employing metagenomics. *PLoS ONE* **15**, e0231210 (2020).
399. Jiang, X. et al. Global meta-analysis of airborne bacterial communities and associations with anthropogenic activities. *Environ. Sci. Technol.* **56**, 9891 (2022).
400. Li, X. et al. A metagenomic-based method to study hospital air dust resistome. *Chem. Eng. J.* **406**, 126854 (2020).
401. King, P. et al. Longitudinal metagenomic analysis of hospital air identifies clinically relevant microbes. *PLoS ONE* **11**, e0160124 (2016).
402. Wu, D. et al. Inhalable antibiotic resistomes emitted from hospitals: metagenomic insights into bacterial hosts, clinical relevance, and environmental risks. *Microbiome* **10**, 19 (2022).
403. Cohen, L. J. et al. Commensal bacteria produce GPCR ligands that mimic human signaling molecules. *Nature* **549**, 48 (2017).
404. Mann, E., Shekarriz, S. & Surette, M. G. Human gut metagenomes encode diverse GH156 sialidases. *Appl. Environ. Microbiol.* **88**, e01755 (2022).
405. Kodzius, R. & Gojobori, T. Marine metagenomics as a source for bioprospecting. *Mar. Genomics* **24**, 21–30 (2015).
406. Puig-Castellví, F. et al. Advances in the integration of metabolomics and metagenomics for human gut microbiome and their clinical applications. *Trends Anal. Chem.* **167**, 117248 (2023).
407. Ariaeenejad, S. et al. Precision enzyme discovery through targeted mining of metagenomic data. *Nat. Prod. Bioprospect.* **14**, 7 (2024).
408. Manara, S. et al. Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol.* **20**, 299 (2019).
409. Shell, W. A. & Rehan, S. M. Comparative metagenomics reveals expanded insights into intra- and interspecific variation among wild bee microbiomes. *Commun. Biol.* **5**, 603 (2022).

Acknowledgements

S.M., C.R., P.J. and N.K.S. are supported by the National Science Foundation (NSF) (grants 2041984 and 2316223) and National Institutes of Health (NIH) (grant R01AI173172). R.C. is supported by ANR Full-RNA, SeqDigger, Inception, PRAIRIE and ERC IndexThePlanet (grants ANR-22-CE45-0007, ANR-19-CE45-0008, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001, EU grant 101088572). This project has received funding through R.C. and F.A. from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreements (No. 872539 and No. 956229). M.P., H.S.M. and T.L. were supported, in part, by the NIH (grant R01AI100947). P.P.L., K.Z. and D.B. are supported by the Polish NCN Sonata BIS (grant number 2020/38/E/NZ2/00598). J.P.Z. and A.M.M. are supported by the NIH (grant U19AI174998). J.P.Z. is supported by the NIH (grant R35GM138369). C.M. thanks I. Tulchinsky and the WorldQuant Foundation, Pershing Square Foundation, Ken Griffin and Citadel, NIH (R01AI125416, R21AI129851, R21EB031466, R01AI151059, U01DA053941, U54AG089334), Rockefeller Foundation and Alfred P. Sloan Foundation (G-2015-13964). D.K., S.L. and J.S.R. are supported by the NIH (grant R01GM146462). E.R.D. is supported by the NIH (grant R35GM146980). E.V.S. was supported by the AFRI Predoctoral Fellowship (grant no. 2022-67011-36461) from the USDA National Institute of Food and Agriculture. V.M., S.M. and A.Z. are supported, in part, by the Ministry of Research, Innovation and Digitization, under Romania's National Recovery and Resilience Plan — funded by the EU — NextGenerationEU program (project no. 760286/27.03.2024, code 167/31.07.2023, within Pillar III, Component C9, Investment 8). V.M. was partially supported by the Government of Republic of Moldova, State Program LIFETECH (No. 020404). A.R. and K.A.C. were supported by the NSF (grant 2109688 to both; grant 2316223 to K.A.C.). E.G. is supported, in part, by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Projects #PEN04752 and #PEN04731 (Accessions #1023328 and #1022444). D.F. and F.S. are supported by the NSF (grant EF-2125142). O.M. and N.M.G. are partly funded by the European Union's Horizon Programme for research and innovation under Grant Agreement No. 101047160 (project BioPIM) and the Swiss National Science Foundation (SNSF).

Author contributions

S.M., S.L., J.S.R., V.M., M.P. and D.K. led the project. S.M. conceived of the presented idea. Introduction (S.M., S.L., J.S.R., V.M., C.R., M.P. and D.K.); Experimentation (S.M., S.L., J.S.R., V.M., A.S.G., N.E.H., S.T., E.V.S., E.G., E.R.D., M.P. and D.K.); Results (S.M., S.L., J.S.R., V.M., C.R., N.K.S., D.B., K.A.C., D.F., A.F., P.J., T.K., P.P.L., W.L., T.L., H.S.M., N.M.G., A.R., F.S., K.Z., M.P. and D.K.); Applications (S.M., S.L., J.S.R., V.M., C.R., R.B., K.D.L., A.M.M., J.P.Z., M.P. and D.K.); Reproducibility and data deposition (S.M., S.L., J.S.R., V.M., M.A., F.A., R.C., P.J., O.M., N.M.G., R.V., A.Z., M.P. and D.K.); Limitations and optimizations (S.M., S.L., J.S.R., V.M., M.P. and D.K.); Outlook (S.M., S.L., J.S.R., V.M., C.M., N.M.G., B.T.T., M.P. and D.K.). All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43586-024-00376-6.

Peer review information *Nature Reviews Methods Primers* thanks Jo Handelsman, Yancong Zhang, Pedro Lebre and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Related links

BBTOOLS: https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/
BlastKOALA: https://www.kegg.jp/blastkoala/
Bracken: https://github.com/jenniferlu717/Bracken
BV-BRC: https://www.bv-brc.org/
Critical Assessment of Massive Data Analysis (CAMDA): https://bipress.boku.ac.at/camda-play/
Critical Assessment of Metagenome Interpretation (CAMI): https://data.cami-challenge.org/
DIAMOND: https://github.com/bbuchfink/diamond?tab=readme-ov-file
ENA: https://www.ebi.ac.uk/ena/
Ensembl: https://ftp.ensembl.org/pub/
FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Galaxy Project: https://galaxyproject.org/cloud/
Human Reference Gut Microbiome (HRGM): https://www.mbiomenet.org/HRGM/HRGM-Genomes.html
HumanMetagenomeDB 1.0: https://web.app.ufz.de/hmgdb/
Joint Genome Institute (JGI) 1000 fungal genomes project (JGI 1K): https://gold.jgi.doe.gov/organisms?Study.GOLD%20Study%20ID=Gs0000001
Kraken: https://ccb.jhu.edu/software/kraken/
KneadData: https://github.com/biobakery/kneaddata
MetaErg: https://github.com/xiaoli-dong/metaerg
MetaSUB Consortium: http://metasub.org/
MGnfy: https://www.ebi.ac.uk/metagenomics/
MG-RAST: https://www.mg-rast.org/

Microbial Genome Database (MBGD): https://mbgd.nibb.ac.jp/
National Microbiome Data Consortium (NMDC): https://microbiomedata.org/
NCBI Cloud Resources: https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/
Picard: https://github.com/broadinstitute/picard
PlanetMicrobe: https://www.planetmicrobe.org/
q2-predict-dysbiosis: https://github.com/bioinf-mcb/q2-predict-dysbiosis
RefSeq: https://ftp.ncbi.nlm.nih.gov/genomes/refseq/
Serratus Database: https://serratus.io/
SRA: https://www.ncbi.nlm.nih.gov/sra
Tara Oceans: http://ocean-microbiome.embl.de/companion.html
TerrestrialMetagenomeDB 2.0: https://web.app.ufz.de/tmdb/
Unified Human Gastrointestinal Genome (UHGG): https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.2/
VEuPathDB: https://amoebadb.org/common/downloads/Current_Release/

[1]Huck Institutes of Life Sciences, Pennsylvania State University, University Park, PA, USA. [2]Department of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Moldova. [3]Department of Biological and Morphofunctional Sciences, College of Medicine and Biological Sciences, Stefan cel Mare University of Suceava, Suceava, Romania. [4]Titus Family Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA, USA. [5]Department of Computer Science, Georgia State University, Atlanta, GA, USA. [6]Sequence Bioinformatics, Department of Computational Biology, Institut Pasteur, Université Paris Cité, Paris, France. [7]Sorbonne Université, Collège doctoral, Paris, France. [8]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA. [9]Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland. [10]Computational Biology Institute, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA. [11]Department of Quantitative and Computational Biology, USC Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA. [12]Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine. [13]Department of Anthropology, Pennsylvania State University, University Park, PA, USA. [14]One Health Microbiome Center, Pennsylvania State University, University Park, PA, USA. [15]Department of Biology, Pennsylvania State University, University Park, PA, USA. [16]Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland. [17]Department of Computer Science, University of Maryland, College Park, MD, USA. [18]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA. [19]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [20]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [21]The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA. [22]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [23]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [24]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [25]Center for Microbial Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [26]Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland. [27]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA. [28]Division of Protective Immunity, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [29]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [30]Department of Animal Science, Pennsylvania State University, University Park, PA, USA. [31]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA. [32]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA. [33]These authors contributed equally: Shaopeng Liu, Judith S. Rodriguez, Viorel Munteanu. [34]These authors jointly supervised this work: Mihai Pop, David Koslicki, Serghei Mangul.