

From sequencing to intelligence: how AI is transforming metagenomics

Fathi A. Mubaraki

Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia

ABSTRACT

Microbial communities are critical in advancing human health. Metagenomics is a technique that analyzes these communities and allows for investigating their composition and functions. Metagenomic shotgun sequencing enables to capture all of the genetic material in environmental samples, such as water, soil, or the human gut. Despite this advantage, one of the main challenges of this technique is the assembling and interpreting of its data, as it produces many short, fragmented reads. Though long-read technologies may change this in the future, artificial intelligence (AI), machine learning (ML) and data science (DS) offer a powerful solution now, enabling scientists to efficiently process and analyze these large and complex datasets. This review explores the latest advancements in AI and ML applications across the metagenomic pipeline. First, it examines the impact of deep learning (DL) on next-generation sequencing, particularly for long-read technologies. Then, it discusses how ML is automating and improving quality control processes, as well as the use of AI applications in metagenome-assembled genome (MAG) assembly, with a focus on contig binning. Finally, this article looks at how AI and ML can improve predictive modeling for phenotype prediction.

Subjects Bioinformatics, Biotechnology, Genomics, Microbiology, Data Mining and Machine Learning

Keywords Metagenomics, Artificial intelligence, Machine learning, Data science, Next-generation sequencing, Shotgun metagenomics, Microbial communities, Deep learning

INTRODUCTION

The Human Microbiome Project studies the various microbial communities that naturally inhabit healthy humans. This project provides a foundational overview of the human microbiome along with its role in health ([Huttenhower et al., 2012](#)). It investigates the healthy human microbiome across different body habitats, such as the gut, skin, and vagina. It has found that the composition of microbial communities varies across healthy individuals. Diet and lifestyle can significantly influence the gut microbiome, and in so doing also contribute to various health outcomes.

Numerous studies have been conducted to determine how the gut microbiome changes with host age. [Tamayo et al. \(2024\)](#) investigated how the composition and diversity of the gut microbiome can affect both healthy and unhealthy aging. They found that diet, physical activity, stress and mental health can positively or negatively impact the gut microbiome and the aging process. [Badal et al. \(2020\)](#) highlighted that older adults, especially centenarians, tend to have higher microbial diversity (both alpha and beta)

Submitted 1 July 2025

Accepted 6 March 2026

Published 4 May 2026

Corresponding author

Fathi A. Mubaraki,
f.mubaraki@ut.edu.sa

Academic editor

Rodolfo García-Contreras

Additional Information and
Declarations can be found on
page 17

DOI 10.7717/peerj.21137

© Copyright
2026 Mubaraki

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

compared to younger adults. The gut microbiota of older adults shows distinct patterns. For example, centenarians (people who live to 100 years of age or beyond) show an enrichment of beneficial microbes, like *Akkermansia*, and a lack of other bacteria, such as *Faecalibacterium* and *Bacteroidaceae*. This diversity is a potential marker of healthy aging and longevity.

Improvements in next-generation sequencing (NGS) technology have enabled the sequencing of microbial communities more comprehensively (both in depth and read length). As a result, metagenomics has become increasingly powerful for understanding these communities. This modern approach enables high-throughput analysis of the collective genome of microbial communities in various environments, including the microbiota in soil, water, the human gut, as well as in understudied environments like the dromedary camel (Mubarak, 2025). In metagenomics, DNA is extracted from a sample which contains multiple microbes, and then sequenced to identify and analyze the entire microbial community at once (Handelsman, 2004; Hugenholtz & Tyson, 2008; Wooley, Godzik & Friedberg, 2010; Thomas, Gilbert & Meyer, 2012). This method helps researchers understand the diversity, functions, and interactions within microbial communities.

Two common techniques used for the study of microbiomes are 16S ribosomal RNA gene sequencing (16S rRNA) and shotgun metagenomics (community WGS). The 16S rRNA approach requires sequencing a conserved gene found across the majority of bacteria (and archaea) to identify and classify microbes within the sequenced community at the genus, or occasionally species, level (Ranjan et al., 2016; Klindworth et al., 2013; Zhang et al., 2023). 16S rRNA gene sequencing is mainly used for taxonomic profiling and is limited to bacteria and archaea. In contrast, shotgun metagenomics sequences all DNA with a sample. This approach covers bacteria and archaea, but extends to all other organisms in the community as well, including viruses and fungi. This approach offers a comprehensive picture of both microbial taxonomy and function, making it ideal for more detailed studies that require functional analysis and/or strain-level resolution (Quince et al., 2017; Sharpton, 2014; Xie et al., 2023).

Figure 1 illustrates the number of studies using 16S rRNA sequencing, compared to the number of studies using shotgun metagenomics, from 2014 to 2023, based on the following search queries:

- (“16S rRNA sequencing”[Title] OR “16S rRNA”[Title]) AND (“2014/01/01”[Date-Publication]: “2023/12/31”[Date-Publication])
- (“shotgun sequencing”[Title] OR “shotgun metagenomics”[Title]) AND (“2014/01/01”[Date-Publication]: “2023/12/31”[Date-Publication])

16S rRNA sequencing was more frequently used throughout this time, and maintained a relatively steady trend, with a slight peak in 2019. Although the use of shotgun metagenomics has grown and continues to do so, it remains less common compared to 16S rRNA sequencing.

In the study of microbial communities, metagenomics is a widely used technique that offers several significant benefits. It enables taxonomic profiling, which identifies and quantifies the abundance of species such as bacteria, fungi, and viruses present in a sample

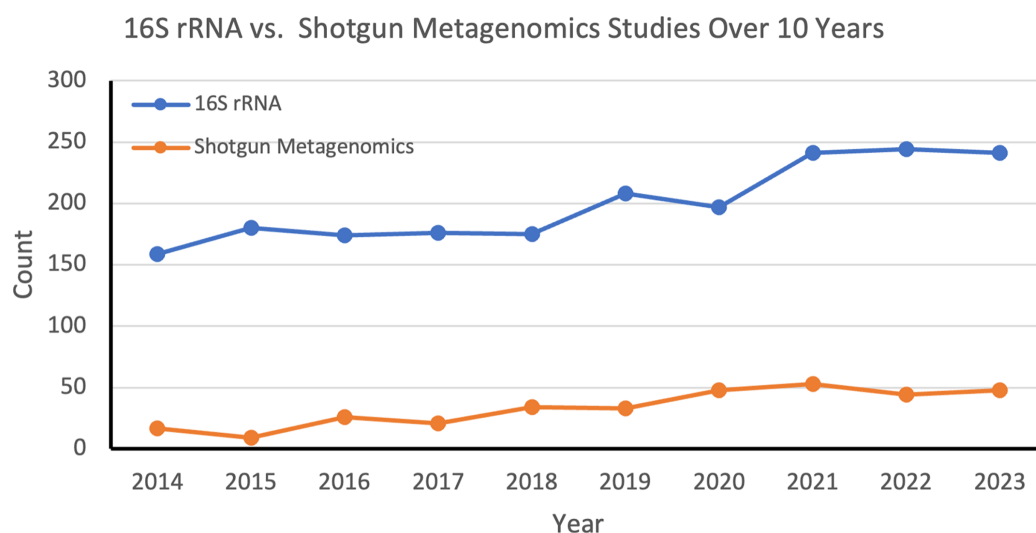


Figure 1 16S rRNA vs. shotgun metagenomics studies over 10 years. The number of studies employing the 16S rRNA and shotgun metagenomics methods, as indexed in the PubMed database, from 2014 to 2023. [Full-size DOI: 10.7717/peerj.21137/fig-1](https://doi.org/10.7717/peerj.21137/fig-1)

(*Shah et al., 2021*), essentially answering the question, “Who is there?” While 16S rRNA sequencing provides lower resolution, shotgun metagenomics delivers deeper coverage and reveals the functional potential of microorganisms through functional profiling. It examines genes in the dataset to answer both “Who is there?” and “What can these organisms do?” (*Hera et al., 2024*).

Artificial intelligence (AI) and machine learning (ML) offer state-of-the-art solutions in metagenomics research. These technologies uncover hidden patterns in complex microbial communities that could be difficult to achieve with traditional univariate statistical approaches. Unsupervised ML, for example, is useful in identifying unknown structures in data without prior knowledge of potential associations. This method has been applied to microbial studies, such as clustering subjects by distinct microbiome subtypes, or partitioning bacterial cell populations based on gene expression profiles, which demonstrates its ability to reveal novel types of information (*Asnicar et al., 2024*).

The review by *Marcos-Zambrano et al. (2021)* provides a comprehensive overview of machine learning algorithms applied to microbiome research. It primarily focuses on methodological aspects of ML models and classification tasks. In contrast, this study focuses on the metagenomic pipeline. It covers AI applications across all stages of metagenomic analysis including sequencing technologies, basecalling, quality control, metagenome-assembled genome, and downstream predictive modeling. Furthermore, this review extends to include recent technologies that were not covered in the earlier study such as third-generation sequencing, AI error correction and real-time microbiome analytics. This review highlights how AI is reshaping the entire metagenomics pipeline. [Table 1](#) summarizes the key differences in scope and focus between previous reviews and the present work.

Table 1 Comparison between previous reviews and the present work.

Aspect	<i>Marcos-Zambrano et al. (2021)</i>	This review
Primary focus	ML algorithms and their applications in microbiome analysis	AI/ML integration across the entire metagenomic pipeline, from sequencing to prediction
Scope	Feature selection, biomarker identification, disease prediction, and treatment	Sequencing technologies, quality control, assembly, functional profiling, and predictive modeling
Data types	Primarily 16S rRNA and shotgun metagenomics	16S rRNA, shotgun metagenomics, long-read sequencing (PacBio (<i>Eid et al., 2009</i>), ONT (<i>Clarke et al., 2009</i>))
Sequencing technologies	Briefly mentioned: focus on NGS short-read data	Extensive coverage of SRS, LRS, hybrid sequencing, and third-generation technologies
Deep learning for basecalling	Not covered	Covered (e.g., DeepConsensus (<i>Baid et al., 2023</i>), MSRCall (<i>Yeh & Lu, 2022</i>), transformer models)
Quality control (QC)	Not a primary focus	Detailed discussion of traditional and AI-based QC for short- and long-read data
MAG assembly	Not a primary focus	Contig assembly, binning (SemiBin2 (<i>Pan, Zhao & Coelho, 2023</i>), VAMB (<i>Nissen et al., 2021</i>)), and quality assessment (CheckM2 (<i>Chklovski et al., 2023</i>))
ML methods coverage	Detailed review of supervised, unsupervised, ensemble methods, deep learning	Overview of ML applications rather than algorithmic comparison

Artificial intelligence has transformed microbiome research. As the field shifts from short-read to long-read sequencing technologies, machine learning has helped improve the accuracy of base-calling in long-read sequencing data, such as Oxford Nanopore and Pacific Biosciences technologies. This advancement ensures higher quality data for downstream analysis. ML also enhances quality control (QC) by automating error detection. Furthermore, AI and ML have advanced metagenome-assembled genome (MAG) assembly by developing more efficient binning and quality assessment tools. They also streamline taxonomic and functional profiling, allowing biologists to quickly identify microbes within a sample, and determine their potential functions. Finally, AI helps build predictive models for disease prediction and proposes personalized treatments based on microbiome data.

Thus, this review explores how these technologies transform each stage of the metagenomics pipeline. [Figure 2](#) illustrates the key steps in this pipeline, and highlights the integration of AI and ML at each stage. In this review, we provide a comprehensive overview of metagenomic analysis that connects traditional metagenomic methods with the emerging applications of AI.

This review is broadly intended for multiple audiences. It is meant to serve as an introduction to modern concepts in metagenomics for new students in the field, in addition to a primer for computational and data scientists to highlight the ways in which microbiologists and microbial ecologists are embracing techniques in machine learning. It also serves as an application note for AI scientists interested in learning how the technologies developed in their field are being applied to a very different one, and as a methods update for established microbiome scientists to showcase new and emerging computational developments in their field which may spur further experimentation and adoption of new AI-enabled methodologies.

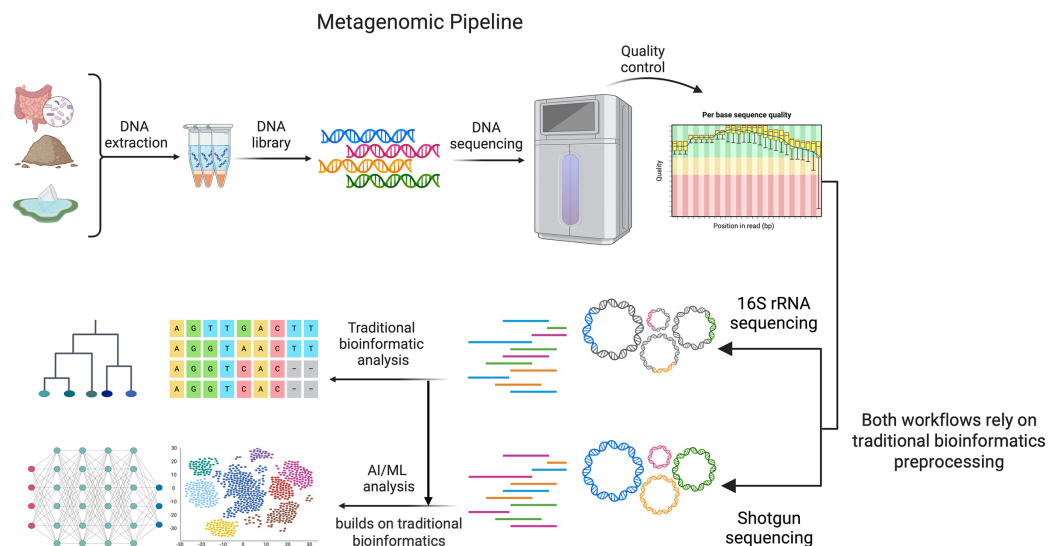


Figure 2 Metagenomic pipeline. The core steps in a metagenomic pipeline are represented here, beginning with initial DNA extraction, then quality control, sequencing (using either traditional 16S rRNA or shotgun metagenomics), and bioinformatics and AI/ML analysis. Although traditional bioinformatic analytical approaches are appropriate for both 16S and shotgun-based metagenomic data, shotgun data provides a much larger feature space (millions of base-pairs per genome *vs* a few hundred, and gene-level annotations). This results in increased potential for deep learning or other advanced AI applications. In both workflows, traditional bioinformatic preprocessing is required, and AI/ML analysis is applied downstream and built on the outputs of these traditional bioinformatic steps. (Created in: BioRender). [Full-size DOI: 10.7717/peerj.21137/fig-2](https://doi.org/10.7717/peerj.21137/fig-2)

SEARCH METHODOLOGY

This review explores the intersection of artificial intelligence (AI), data science, and metagenomics. A literature search was conducted using PubMed and Google Scholar in order to identify the studies included in this review. Keywords including machine learning, artificial intelligence, long read, short read, shotgun metagenomics, next-generation, third-generation sequencing, and microbiome were used (530 initial combined search results). Peer-reviewed articles that were published from 2004 to 2024 were focused on, with particular emphasis on recent publications. Non-English articles were excluded from the search. To ensure comprehensive coverage of the field, earlier but highly cited studies were also included when relevant. In total, 73 references were selected for inclusion in this review.

METAGENOMIC SEQUENCING METHODS

Metagenomic sequencing can be divided into short-read, long-read, and hybrid sequencing. Short-read sequencing (SRS), sometimes referred to as next generation sequencing (NGS), has accelerated genetic research. It enables high-throughput analysis at a lower cost compared to traditional Sanger sequencing. Characterized by its ability to massively parallel sequencing of short DNA fragments, typically ranging from 50 to 300 base pairs (Hu *et al.*, 2021; Kumar, Cowley & Davis, 2024; McCombie, McPherson & Mardis, 2019), SRS can generate millions to billions of reads per run. This technology has greatly increased sequencing output. It enabled researchers to tackle complex biological

Table 2 Comparison of SRS and LRS technologies.

	Short-read sequencing	Long-read sequencing
Throughput	High throughput	Lower throughput
Read length	(~50–300 bp)	10 kb+
Cost	Lower cost	Higher cost
Platforms	Illumina, MGI, Ion Torrent	PacBio SMRT, ONT

questions involving whole genomes and metagenomes (Jeon *et al.*, 2021). Platforms such as Illumina, MGI, and Ion Torrent are the most commonly used for SRS. Each platform provides unique advantages in terms of throughput, read length, accuracy, and cost-effectiveness (Kumar, Cowley & Davis, 2024; van Dijk *et al.*, 2014; Quail *et al.*, 2012). The high throughput, shorter read lengths, and cost advantages of SRS, as highlighted in Table 2, make it accessible for many genomic applications.

Although SRS has significantly advanced the microbiome field, it faces limitations due to its relatively short read lengths. This raises difficulties in resolving repetitive regions and assembling complex genomes accurately (Kumar, Cowley & Davis, 2024; van Dijk *et al.*, 2014). In the context of repetitive regions, the lack of read-level specificity makes it difficult to assign many short reads to a single organism. This problem becomes more obvious when similar species are present in the sample. As a result, more sophisticated bioinformatics algorithms are necessitated by SRS to make sense of the large but comparatively disjointed data generated.

Long-read sequencing (LRS) is sometimes called “third-generation sequencing” (TGS). It allows for the analysis of DNA and RNA molecules in greater detail than short-read methods. Unlike SRS, which relies on the fragmentation of DNA, LRS technologies directly sequence native DNA or RNA molecules. They generate reads that can span tens of thousands of bases and even exceed a million bases in some cases (Amarasinghe *et al.*, 2020; De Coster, Weissensteiner & Sedlazeck, 2021; Logsdon, Vollger & Eichler, 2020). Examples of this method include Pacific Biosciences’ (PacBio) single-molecule real-time (SMRT) sequencing (Eid *et al.*, 2009) and Oxford Nanopore Technologies’ (ONT) nanopore sequencing (Clarke *et al.*, 2009).

As illustrated in Table 2, LRS techniques provide longer read lengths. This ability to sequence long fragments provides distinct advantages over SRS approaches. These include improved *de novo* assembly, more accurate mapping, enhanced identification of transcript isoforms, and increased detection of complex structural variants (Amarasinghe *et al.*, 2020; Satam *et al.*, 2023), which utilize shorter read lengths. Notably, the longer reads produced by LRS are particularly useful for accurately assigning each read to a specific reference organism. They are also valuable for resolving repetitive genomic regions and complex structural variations. In addition, long-read sequencing enables strain-level resolution. It allows researchers to distinguish closely related microbial strains that short reads could not (Afolayan *et al.*, 2024). All of these create challenges for short-read taxonomy assignment and assembly methods.

Despite these advantages, LRS technologies have a lower throughput compared to SRS, as well as higher error rates ([Amarasinghe et al., 2020](#); [Warburton & Sebra, 2023](#); [Conlin et al., 2022](#)). LRS platforms often produce fewer reads per unit of time than SRS platforms. Therefore, despite the longer reads afforded by LRS, the total amount of data generated in a single sequencing run may be less than a comparably priced SRS run. Long-read sequencing platforms can generate reads with lengths spanning from tens to hundreds of kilobases, but the longer the read, the more errors it is likely to contain. This is especially true if there are repetitive or homopolymer regions ([Tørresen et al., 2019](#); [Zhang, Jain & Aluru, 2020](#)). Another disadvantage is cost. Though the cost of long-read sequencing has significantly decreased over the years, it remains higher than SRS. For example, the cost of PacBio Sequel II HiFi and ONT PromethION remains relatively high. For instance, at comparable sequencing quality to SRS, PacBio Sequel II HiFi sequencing offers highly accurate reads with over 99% accuracy, and generates 25-fold coverage of a human genome. However, this process requires three SMRT Cell 8M which increases the cost by two to three times that required for continuous long-read (CLR) data, which has lower single-pass read accuracies, typically in the 85–90% accuracy range ([Logsdon, Vollger & Eichler, 2020](#)).

Recent advances in Oxford Nanopore Technologies, specifically their Kit 14 chemistry, have greatly improved sequencing performance. In tandem with AI-enabled base-calling, these developments are beginning to enable simplex (single molecule, without consensus) accuracy of over 99% per read. This represents an exciting leap in LRS usability. Newer AI models, including a new basecaller using a transformer architecture, are pushing these accuracies even higher with existing chemistries ([Oxford Nanopore Technologies, 2024](#)).

[Figure 3](#) illustrates sequencing platform usage in NCBI's Sequence Read Archive (SRA) ([Leinonen et al., 2011](#)), retrieved on December 21, 2024. It shows Illumina's dominance, with NovaSeq and HiSeq contributing the most records due to their high throughput and cost efficiency. In contrast, LRS platforms like Oxford Nanopore and PacBio SMRT have fewer records. However, they are essential for applications requiring longer reads, such as *de novo* assembly and structural variation analysis. Their lower usage reflects higher costs and lower throughput.

Hybrid sequencing combines data from both long-read and short-read DNA sequencing technologies to leverage the strengths of each. This approach aims to produce more accurate and complete genomic assemblies ([De Maio et al., 2019](#)). Hybrid assembly may be more advantageous than long-read-only assemblies for complex applications like metagenomics and single-cell genomics. One such tool that demonstrates the power of hybrid assembly is Unicycler, a hybrid assembly tool designed specifically for bacterial isolated genomes. It combines the accuracy of SRS with the structural resolve of LRS to generate high-quality assemblies ([Wick et al., 2017](#)). Unicycler begins by constructing an accurate and connected assembly graph from short reads using SPAdes. Then it refines this graph by aligning long reads to identify the best paths. This approach is also reflected in algorithms like HybridSPAdes, which combine the precision of short-read from platforms like Illumina with the coverage provided by long-read such as SMRT and nanopore

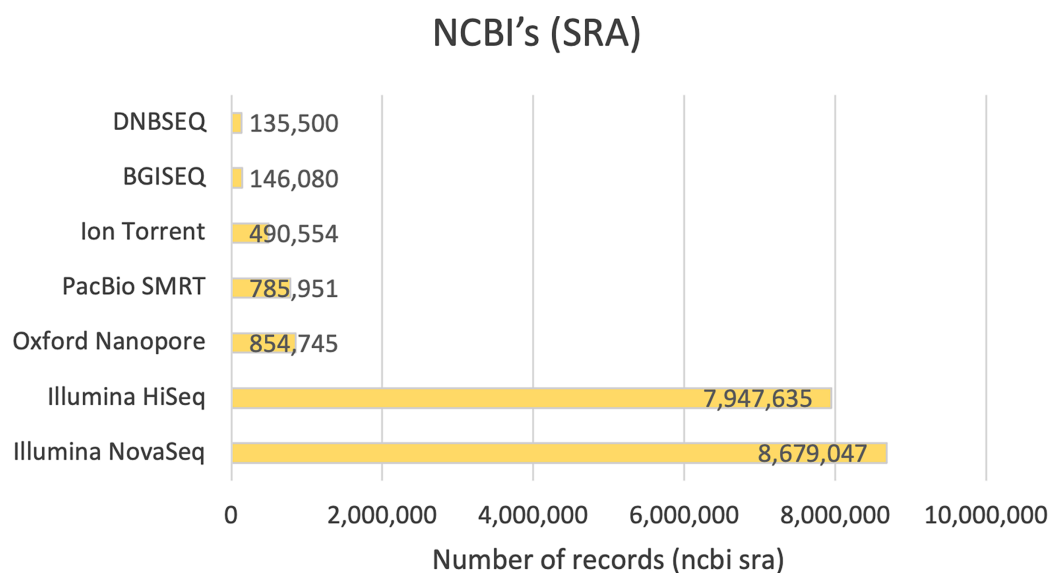


Figure 3 Number of records (NCBI SRA). The usage of various sequencing platforms based on the number of records in NCBI's SRA, with SRS platforms (HiSeq, NovaSeq, Ion Torrent, BGISEQ, and DNBSEQ) accounting for approximately 92.04% of the total records.

Full-size  DOI: [10.7717/peerj.21137/fig-3](https://doi.org/10.7717/peerj.21137/fig-3)

sequencing. The hybrid method efficiently assembles genomes even with low coverage of long reads, thereby significantly reducing sequencing costs (*Antipov et al., 2016*).

DEEP LEARNING FOR BASE-CALLING

Oxford Nanopore and Pacific Biosciences (ONT and PacBio) have both recently integrated deep learning (DL) techniques to improve the accuracy of base-calling during DNA sequencing. These technologies use deep learning in order to better differentiate between nucleotide possibilities, even when the raw signals might be weak or ambiguous. A recent study by *Baid et al. (2023)* presented a new method called DeepConsensus, which uses advanced deep learning models to improve the accuracy of DNA sequencing with Pacific Biosciences' technology. DeepConsensus improves on older methods like pbccs by reducing sequencing errors. The sequences it reads are more accurate, making it easier to assemble a complete and correct genome (*Baid et al., 2023*). For instance, when compared to the traditional pbccs, DeepConsensus significantly reduced read errors by 42%. This improvement increased the quality of the HiFi reads across different quality levels: Q20 by 9%, Q30 by 27%, and Q40 by 90%.

DeepConsensus uses a transformer–encoder, a type of deep learning (DL) model, to correct mistakes it reads in the DNA sequences. This method is shown to increase the quality and quantity of accurate reads, making the sequencing data more reliable. As a result, it improves the genome assembly—boosting the NG50 contiguity from 4.9 megabases (Mb) to 17.2 Mb and increasing gene completeness from 94% to 97%. It also reduces the false gene duplication rate from 1.1% to 0.5%. In addition, it improves assembly base accuracy from Q43 to Q45, and decreases variant-calling errors by 24%

([Baid et al., 2023](#)). Furthermore, because DeepConsensus can accurately handle longer DNA sequences, it enables scientists to study more complex genetic areas, such as metagenomics and biodiversity within bacterial communities. The study suggested that the methods developed for DeepConsensus could be adapted for other research tasks, potentially revolutionizing the study of genetic variation and interactions within and between species.

Another article introduced MSRCall, a new deep learning model designed to improve how the MinION device reads DNA sequences ([Yeh & Lu, 2022](#)). The MinION, made by Oxford Nanopore Technologies, is notable for its portability and ability to deliver long DNA readouts quickly ([Jain et al., 2016](#)). However, it faces challenges with accuracy, because the signals it reads are complex and noisy. The MSRCall model uses DL techniques, including a multi-scale structure with convolutional and recurrent layers. It combines different layers that analyze the signals at multiple scales, which helps in recognizing both immediate and distant patterns in the DNA strands. This approach allows MSRCall to be more accurate than previous methods. For example, during testing, MSRCall demonstrated better performance compared to the traditional SACall method ([Huang et al., 2022](#)). It increased read accuracy from approximately 91% to 92.215% and boosted consensus accuracy from around 99.719% to 99.952%.

At a recent community meeting in Houston, Texas, Oxford Nanopore Technologies announced major improvements in DNA and RNA sequencing. They achieved a new high in sequencing accuracy, with results showing 99.8% accuracy for individual DNA molecules (single sequencing reads). The longest read reached 99.9% accuracy over 1.1 Mb ([Oxford Nanopore Technologies, 2023](#)) using a deep learning AI model based on the transformer architecture.

Building on these advancements, the integration of AI and DL has transformed sequencing technologies, especially in metagenomic studies. Microbial ecology and microbiome study greatly benefit from the increased accuracy and information content that AI methods like DeepConsensus can provide. These advances can in turn help in identifying novel microbes, tracking microbial evolution, studying microbial interactions in their natural environments, and more, including the recovery of low-abundance genomes (in assembly) and strain-level resolution (in profiling), thereby improving understanding of the microbiota.

METAGENOMIC PIPELINE

Quality control

Traditional QC (short reads)

Quality control (QC) is an important early step in the sequencing process, as it ensures that the data at the beginning of the analysis funnel is sufficiently accurate or reliable before using it for further analysis downstream. Although many tools have been developed to implement or improve upon QC for sequencing, most of these tools are intended for use with short-read sequencing (SRS) and are not well-suited for third-generation (long-read, LRS) sequencing. There is hence a need for new or improved QC tools.

To illustrate this limitation, several existing tools designed for NGS can be considered. The NGS QC Toolkit (*Patel & Jain, 2012*) offers comprehensive quality control for sequencing data, while FaQCs (*Lo & Chain, 2014*) provides rapid evaluation and QC. Other tools, like MultiQC (*Ewels et al., 2016*), which aggregates and summarizes analysis results from multiple tools and samples into a single report, can help streamline the QC process. Some studies have also demonstrated that effective QC can be achieved even in the absence of a reference genome (*Trivedi et al., 2014*). Despite their usefulness in certain situations, these tools must be adapted to support third-generation sequencing to remain applicable.

Traditional QC (long reads)

A study by *Fukasawa et al. (2020)* introduced LongQC as an innovative QC tool specifically optimized for third-generation sequencing such as Oxford Nanopore and SMRT sequencing from Pacific Bioscience. This tool excels in automated and efficient QC, covering all major TGS platforms without the need for reference genomes. Similarly, the enhancements to SAMStat 2, as detailed by *Lassmann (2023)*, provided significant improvements, including support for paired-end and long-read data. This expands its use to newer sequencing technologies, allowing the tool to handle more data types and improving its usefulness in sequencing QC. However, LongQC relies on k-mer based approaches. While these are computationally efficient, they might miss the errors that DL models could identify by learning from complex patterns in the data. Similarly, SAMStat 2 still primarily focuses on extracting and visualizing quality metrics rather than predictive models.

AI for short-reads QC

Al-Ghalith et al. (2018) published SHI7, a QC pipeline initially optimized for SRS. Notably, this pipeline automatically adjusts parameters used by the QC process based on quality parameters inferred from the sequencing data itself, enabling support for datasets from different sequencing technologies without requiring extensive user input. SHI7's learning mode autonomously determines the best QC parameters for each dataset. It makes decisions on adapter trimming, quality and length filtering, and whether it is feasible to stitch reads together. Recently, SHI7 was expanded to be compatible with long-read sequences, broadening its utility across sequencing platforms (SHI7 GitHub repository; available at <https://github.com/knights-lab/shi7>, accessed 7 April 2026).

Albrecht et al. (2021) developed SeqQscorer, a software tool that applies ML to enhance QC processes for NGS data. Using tree-based and DL algorithms, SeqQscorer automates QC by evaluating NGS data quality, thereby improving efficiency. In their study, the authors analyzed 2,642 NGS files from the ENCODE data portal and systematically benchmarked ML algorithms to construct precise, unbiased predictive models. These models, validated across diverse species, demonstrate SeqQscorer's accuracy. Additionally, Jiajin Li and colleagues introduced ForestQC (*Li et al., 2019*), a tool that employs Random Forest models (*Breiman, 2001*) to improve genetic data QC. It evaluates sequencing quality

indicators, like depth and genotyping quality, to distinguish between high- and low-quality variants.

AI for long-read QC

Herro is a deep learning model designed to correct errors in Oxford Nanopore Technologies (ONT) simplex ultra-long sequencing reads ([Stanojević et al., 2024](#)). It enables telomere-to-telomere (T2T)-phased genome assemblies without the need for multiple sequencing technologies. Herro can improve accuracy by up to 100-fold. It supports both R9.4.1 and R10.4.1 ONT Simplex reads and integrates with *de novo* assemblers like Hifiasm and Verkko. To ensure a high-quality final genome, it utilizes consensus polishing, which serves as its final step. In this process, a draft assembly is optimized using all reads to correct the reference assembly rather than individual reads. This step builds greater confidence in the final result. Tools like Medaka ([Oxford Nanopore Technologies, 2025b](#)), which now includes a neural network consensus caller for bacteria and metagenomes, and Dorado Polish ([Oxford Nanopore Technologies, 2025a](#)), recently released by ONT, use advanced neural networks to achieve highly accurate polished assemblies.

MAG assembly: contig, binning, and quality assessment with AI

Reconstructing individual microbial genomes from metagenomic data is a crucial step, especially when the reference databases are limited, or when strain-level and/or genome-resolved functional profiling is desirable. Contigs are continuous DNA sequences assembled from shorter overlapping reads. These contigs are grouped into bins through a process called binning ([Du & Sun, 2022](#)). This can be challenging due to the presence of repetitive, and highly conserved regions as well as related organisms within the same sample. Traditionally, for both short and long-read datasets, tools like metabat 2 ([Kang et al., 2019](#)) were used. These tools use clustering approaches to group contigs (partially assembled DNA that are not complete genomes) by their tetranucleotide frequency and abundance. AI and ML are powerful tools that may greatly help with this task.

Applying machine learning techniques helps improve the reconstruction of individual genomes from complex metagenomic data. In a recent study, [Pan, Zhao & Coelho \(2023\)](#) introduced SemiBin2, a self-supervised learning tool for both short- and long-read datasets. SemiBin2 achieves high binning accuracy while using less time and memory than SemiBin1. It also performs better than VAMB, a deep learning-based metagenomic binning tool ([Nissen et al., 2021](#)). For long-read datasets, it can reconstruct nearly complete bins. These advances make SemiBin2 one of the most reliable tools for metagenomic analysis.

Despite these advances, even state-of-the-art binning algorithms have their limits. Evaluating the quality of metagenomic assembled genomes (MAGs) after the contig assembly and binning processes is essential. This quality assessment estimates the completeness and contamination of each bin or MAG (as a percentage). It allows downstream QC to ensure that the genomes are complete and free from contamination to

preestablished thresholds (such as the MiMAG criteria) (Bowers *et al.*, 2017). This process helps confirm that the assembled genome data are more reliable for further analysis.

CheckM2 (Chklovski *et al.*, 2023) is a machine learning tool built to assess the quality of microbial genomes, especially MAGs. It can predict bacterial and archaeal genome completeness and contamination without relying on taxonomic information. CheckM2 uses artificial neural networks (NNs) and gradient boosted (GB) decision trees, which can quickly and accurately evaluate MAG quality. These models are trained on simulated genomes (synthetic MAGs) generated from known prokaryotic reference genomes to predict completeness and contamination, based on KEGG-annotated protein features (Kanehisa & Goto, 2000). It also reports features like genome length and number of coding sequences. The authors showed that CheckM2 performs well in predicting completeness and contamination of genome quality.

AI AND MACHINE LEARNING IN MICROBIOME STUDIES: EXPLORATION AND PREDICTION

AI/ML for visualizing microbiome data

The field of metagenomic research deals with complex, high-dimensional datasets. These datasets can contain thousands of different bacterial species along with their gene annotation patterns. Traditional methods like principal component analysis (PCA) (Pearson, 2010) can struggle with high-dimensional compositional data because, as a linear method, they fail to effectively capture non-linear relationships.

In contrast, DNABERT-S (Zhou *et al.*, 2025) is a deep learning model developed to classify and group DNA sequences by species, even when reference genomes are not available. It converts DNA sequences into numerical representations, allowing the grouping of sequences from the same species in a shared space. This process helps the model identify unique sequence patterns that distinguish DNA from different species. The model uses two key techniques. Manifold Instance Mixup combines hidden features to generate more challenging training examples, while Curriculum Contrastive Learning trains the model iteratively from simple tasks to progressively more complicated ones. DNABERT-S, per the authors, performs well at species clustering, classification, metagenomics binning, and other DNA discrimination tasks, which makes it a useful tool for analyzing complex DNA data.

DNABERT-S models also serve as a proof of concept as to how AI can address long-standing problems in metagenomics, from species-specific DNA foundational models to binning and classification. AI methods provide a means to advance the state of the art beyond traditional methods and make it easier to study microbial communities more accurately and efficiently.

AI/ML for phenotype prediction

Predictive modeling in microbiome research involves using computational techniques such as ML to model, estimate, or predict outcomes of biological systems. It is arguably the most critical component of an analysis because it helps identify relationships between microbes and diseases. It also feeds the field's understanding of microbial roles in the

environment. [Hernández Medina et al. \(2022\)](#) highlighted how ML techniques, like Random Forests (RFs) and regression models, are applied to microbial classification, and phenotype prediction. They also showed how recurrent neural networks (RNNs) have been applied to predict changes in microbial communities over time and their association with specific phenotypes. To illustrate this, they referenced the work of [Metwally et al. \(2019\)](#), who created a predictive model based on RNNs that analyzed longitudinal microbiome profiles to predict food allergies in infants over a three-year period.

Beyond association research, a language model has also been developed for the human gut microbiome using natural language processing techniques to capture microbial interactions and patterns. This self-supervised model improves prediction accuracy for conditions like Irritable Bowel Disease (IBD). It achieves this by using deeply embedded hidden features that emerge from the foundation model's deep understanding of the structure of the human microbiome. This model works well with different datasets and shows how latent features of microbial communities can affect health ([Pope et al., 2023](#)). In the opposite vein, a study by [Gou et al. \(2020\)](#) investigated the relationship between gut microbiome features and type 2 diabetes risk using an interpretable ML framework across multiple Chinese cohorts. Its emphasis on interpretable and simplified features distinguishes it from much of the deep learning "black box" models. Instead of relying on hidden representations, it explicitly identifies sets of simpler features with predictive power in a "score." The microbiome risk score (MRS) correlates strongly with diabetes risk and glucose levels. This study found that body fat distribution and other demographic factors influence the microbiome-diabetes relationship.

[Walsh, Stallard-Olivera & Fierer \(2023\)](#) presented a practical guide to using ML in microbial ecology that emphasizes the importance of exploring and preprocessing microbiome data before modeling. These considerations align with the findings of [Marcos-Zambrano et al. \(2021\)](#), whose study provided a comprehensive review of the most frequently used ML methods in human microbiome research. These include supervised (logistic regression, k-nearest neighbors, naive Bayes, support vector machines) and unsupervised methods such as (hierarchical clustering, biclustering, non-negative matrix factorization). For instance, Random Forests (RFs) are used extensively for classification and biomarker identification due to their ability to handle high dimensionality data. The review mentioned their use in various disease predictions and highlighted their strong performance.

While machine learning models often perform well in prediction tasks, interpretable methods like Shapley Additive Explanations (SHAP) ([Lundberg & Lee, 2017](#)) or feature importance rankings from RFs are crucial for gaining biological insights and ensuring the results are meaningful ([Walsh, Stallard-Olivera & Fierer, 2023](#)). [Marcos-Zambrano et al. \(2021\)](#) highlighted examples of deep learning applications in microbiome research. One example is the work of [Asgari et al. \(2018\)](#), who used DL methods along with Random Forest and Support Vector Machine (SVM) ([Cortes & Vapnik, 1995](#)) to classify human body sites and diagnose Crohn's disease with high performance in large datasets.

Similarly, they described a study by [Fioravanti et al. \(2018\)](#) that introduced a phylogenetic Convolutional Neural Network (CNN) ([Lecun et al., 1998](#)) for classifying gut

microbiome data. This model categorized data into six phenotypes. The dataset consisted of 38 controls and 222 IBD patients. Data were pre-processed using QIIME 2 ([Bolyen et al., 2019](#)), UCLUST ([Edgar, 2010](#)), and RaxML ([Stamatakis, 2014](#)) to calculate relative abundance, cluster taxa, and construct a phylogenetic tree as input for the CNN. The CNN outperformed traditional models such as Linear Support Vector Machine (LSVM) and Random Forest in all tasks. This further demonstrates creative uses of ML and AI tools to improve performance in downstream microbiome classification tasks.

AI/ML for predicting microbial functions

Several studies have applied AI methods to multi-omics data for functional profiling. [Wu et al. \(2024\)](#) showed how metatranscriptomics, metagenomics, and glycomics can be combined with AI to predict diseases. An individual's gut microbiome can be used as biomarkers for personalized treatment predictions. In this case, machine learning can be applied to predict patterns in a patient and identify microbial biomarkers, which may help improve individual treatments.

A recent study by [Odrzywolek et al. \(2022\)](#) introduced a deep learning model designed to study microbial proteins, particularly when traditional methods fall short due to the novelty and complexity of the data. The model was built using Bidirectional Long Short-Term Memory (BiLSTM) networks ([Thireou & Reczko, 2007](#)) and trained on a large database to predict the functions of microbial proteins, especially those absent from existing databases. To validate the model, the researchers tested it on the SwissProt database, a well-established resource for protein information ([Bairoch & Apweiler, 2000](#)). Their results showed that the model effectively identifies features related to protein structure and function.

Deep learning is also being used to predict protein structures. One major example is AlphaFold2, developed by DeepMind, which can predict a protein's 3D shape from its amino acid sequence. New versions like AlphaFold-Multimer and ESMFold can analyze protein complexes and unknown microbial proteins ([Levy Karin & Steinegger, 2025](#)). These tools have been applied in several advanced use cases, such as detecting similarity despite low pairwise sequence identity, bypassing molecular replacement, and identifying potential protein interactors. By applying these AI tools in metagenomic studies, researchers can better understand the function of proteins more deeply. For example, they can uncover hidden microbial enzymes that may play roles in metabolism or disease.

Microbiome metabolomics is the study of the metabolites, or the small molecules produced by cellular metabolic processes ([Bauermeister et al., 2022](#)). These can be powerful markers for the downstream activity of microbes and provide a window into the chemistry and bioactivity of a community. They have profound implications in understanding the mechanistic link between microbes and their environment. However, metabolomics is a challenging and expensive modality that is rarely performed on stool samples, thus presenting numerous challenges for analysis. Despite these challenges, metabolomics techniques go well beyond identifying microbial species, and enable the inference of metabolic activities and functional roles.

Table 3 Applications of AI/ML across different stages of the metagenomic studies.

Stage	AI/ML application	Example Tools/Techniques
Sequencing	Improving base-calling accuracy for long-read technologies	DeepConsensus (Baid et al., 2023), MSRCall (Yeh & Lu, 2022)
Quality control	Automating error detection and improving QC processes	LongQC (Fukasawa et al., 2020), SHI7 (Al-Ghalith et al., 2018), SeqQscorer (Albrecht et al., 2021)
Assembly	Enhancing contig binning and MAG assembly	MetaBAT 2 (Kang et al., 2019), VAMB (Nissen et al., 2021)
Taxonomic profiling	Identifying microbial species and their abundance	TIPP2 (Shah et al., 2021), DNABERT-S (Zhou et al., 2025)
Functional profiling	Predicting microbial functions and metabolic pathways	mNODE (Wang et al., 2023), PICRUST (Langille et al., 2013)

[Wang et al. \(2023\)](#) introduced a method called mNODE to predict metabolomic profiles from microbial community data. This approach helped address the challenge of expensive and complex metabolomics experiments by using a ML technique called Neural Ordinary Differential Equations (NODE) ([Chen et al., 2019](#)), which is built on deep neural network models, to computationally predict the metabolomic profiles of microbiomes directly from metagenomic sequencing data. This method was tested on both synthetic and real-world data and shows better accuracy than existing methods. This approach can theoretically augment existing ubiquitous metagenomic sequencing data with AI predictions of its metabolites. As a result, it may reduce the need for expensive metabolomic experiments in many cases. It can also improve the value of older datasets, and accelerate microbiome research with fast, cost-effective metabolic predictions. This echoes the approach taken by PiCRUST ([Langille et al., 2013](#)) for functional predictions (using Bayesian ancestry reconstruction) using only limited marker gene (16S) compositional data. It is an example of predictive modality augmentation, which uses one type of (limited) data to generate first-order approximations of a different or more complex downstream data type.

[Table 3](#) provides an overview of AI/ML applications across different stages of the metagenomic pipeline, including sequencing, quality control, assembly, and functional profiling. These advancements highlight the role of AI and ML.

DISCUSSION

AI and ML have already transformed metagenomic research by addressing key challenges such as data complexity, QC, feature generation (through better assemblies and new foundation models), and downstream predictions of outcomes (or even entire data modalities). AI and ML can advance nearly every major stage of routine microbiome analysis, from data acquisition, to QC, to predictive modeling. To the latter point, AI and ML have demonstrated the potential to find patterns in complex microbial communities, enabling more accurate predictions about microbial functions, disease association, and environmental interactions.

As AI research advances and becomes more integrated into microbiome analysis, it is anticipated that more breakthroughs in metagenomics and microbiome research will be attributable to it. One such possibility for AI to make substantive improvements is the

development of real-time microbiome analytics. With the advent of portable sequencing technologies like Oxford Nanopore's MinION, combined with AI-driven base-calling and error correction tools, it is possible to perform real-time sequencing and analysis of microbial communities in the field. For example, real-time microbiome analysis could enable the identification of pathogens during a disease outbreak. It could also provide immediate feedback on the impact of dietary interventions on gut microbiota. These could both improve patients' lives.

Another area that AI has the potential to improve is the integration of AI with multi-omics data. It is often desirable to combine metagenomics data with other omics types, such as transcriptomics, proteomics, metabolomics, and epigenomics. Doing so can allow for a more comprehensive systems-level understanding of microbial ecosystems (often in tandem with host biology and 'omics). AI models are becoming able to natively integrate and analyze multi-omics data, and will thus be essential for more clearly studying the complex relationships between microbial communities and their hosts ([Gupta et al., 2024](#)).

However, it deserves to be mentioned that as AI and ML continue to be developed and become increasingly more capable, it is important to address their various challenges, particularly related to interpretability of the models and their inference process. It is often difficult to understand how a "black box" deep learning model arrives at its predictions. Additionally, as models become increasingly complex and are brought to feature spaces where the number of features far outstrips the sample size (the " $p \gg N$ " regime), overfitting becomes another concern. Thus, ensuring the interpretability and generalizability of any types of models will be critical for their widespread adoption.

While the main focus of this review is to discuss the latest advancements in AI across the metagenomic pipeline, it is important to also consider the ethical implications and data reproducibility of AI based biological research. [Asnicar et al. \(2024\)](#) highlight the principle of "garbage in, garbage out," which means the results of a data processing pipeline, including AI/ML-based pipelines, are sensitive to input data quality. Models trained on datasets that have errors and noise are likely to produce misleading and unreliable results. In other words, the performance of machine learning models strongly depends on the quality of the data that are used for training. Additionally, both public metagenomic databases and AI bioinformatics tools are updated frequently. This can lead to different results when the same analyses are repeated at different times. For example, perfectly reproducing a study conducted a decade ago using R version 2.x can be difficult because it often requires downgrading software versions ([Ziemann, Poulain & Bora, 2023](#)), especially if packages used at the time have since been removed. To reduce data bias, studies should clearly report data source and preprocessing steps. Reproducibility can be also improved by using version-controlled workflows and documenting all software versions used throughout the analysis (from raw data QC to final figure generation). If randomization is used, a random seed should be set where possible to ensure reproducibility of results within a given software version. It is the scientific and ethical responsibility of researchers to ensure repeatable, reproducible results and transparency regarding the tools used.

CONCLUSIONS

This review studies recent advances in artificial intelligence and machine learning across the metagenomic analysis pipeline. It highlights how these methods can handle complex data and reveal new biological insights. AI and ML are increasingly being integrated into the toolset used for analyzing microbiome data, and depending on the application currently represent both iterative refinements to existing tooling as well as sharper changes in capabilities. From slight increases to base quality or QC, to future real-time microbiome analyses and seamless integration of noisy multi-omics data, AI has the potential to drive both gradual evolution and breakthroughs in multiple fields, ultimately improving researchers' understanding of human microbiota.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The author received no funding for this work.

Competing Interests

The author declares that they have no competing interests.

Author Contributions

- Fathi A. Mubaraki conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:
This is a literature review.

REFERENCES

- Afolayan AO, Montaña SA, Akintayo IJ, dos SLD, Reuter S. 2024. Assessing the performance of current strain resolution tools on long-read metagenomes 2024. Available at <https://doi.org/10.1101/2024.11.20.624313>.
- Al-Ghalith GA, Hillmann B, Ang K, Shields-Cutler R, Knights D. 2018. SHI7 is a self-learning pipeline for multipurpose short-read DNA quality control. *mSystems* 3:e00202-17 DOI 10.1128/msystems.00202-17.
- Albrecht S, Sprang M, Andrade-Navarro MA, Fontaine J-F. 2021. seqQscorer: automated quality control of next-generation sequencing data using machine learning. *Genome Biology* 22(1):75 DOI 10.1186/s13059-021-02294-2.
- Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21(1):30 DOI 10.1186/s13059-020-1935-5.
- Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32(7):1009–1015 DOI 10.1093/bioinformatics/btv688.

- Asgari E, Garakani K, McHardy AC, Mofrad MRK. 2018. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics* 34(13):i32–i42 DOI 10.1093/bioinformatics/bty296.
- Asnicar F, Thomas AM, Passerini A, Waldron L, Segata N. 2024. Machine learning for microbiologists. *Nature Reviews Microbiology* 22(4):191–205 DOI 10.1038/s41579-023-00984-1.
- Badal VD, Vaccariello ED, Murray ER, Yu KE, Knight R, Jeste DV, Nguyen TT. 2020. The gut microbiome, aging, and longevity: a systematic review. *Nutrients* 12(12):3759 DOI 10.3390/nu12123759.
- Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, Yang H, Kolesnikov A, Ammar W, Vert J-P, Vaswani A, McLean CY, Nattestad M, Chang P-C, Carroll A. 2023. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nature Biotechnology* 41:232–238 DOI 10.1038/s41587-022-01435-7.
- Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28(1):45–48 DOI 10.1093/nar/28.1.45.
- Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. 2022. Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews Microbiology* 20(3):143–160 DOI 10.1038/s41579-021-00621-9.
- Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolk T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Priesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37(8):852–857 DOI 10.1038/s41587-019-0209-9.
- Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloie-Fadrosch EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Murat Eren A, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35(8):725–731 DOI 10.1038/nbt.3893.
- Breiman L. 2001. Random forests. *Machine Learning* 45(1):5–32 DOI 10.1023/A:1010933404324.

- Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. 2019. Neural ordinary differential equations. Available at <https://doi.org/10.48550/arXiv.1806.07366>.
- Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods* 20(8):1203–1212 DOI 10.1038/s41592-023-01940-w.
- Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* 4(4):265–270 DOI 10.1038/nnano.2009.12.
- Conlin LK, Aref-Eshghi E, McEldrew DA, Luo M, Rajagopalan R. 2022. Long-read sequencing for molecular diagnostics in constitutional genetic disorders. *Human Mutation* 43(11):1531–1544 DOI 10.1002/humu.24465.
- Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297 DOI 10.1007/BF00994018.
- De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nature Reviews Genetics* 22(9):572–587 DOI 10.1038/s41576-021-00367-3.
- De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Walker AS, Stoesser N, on behalf of the REHAB consortium. 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics* 5(9):e000294 DOI 10.1099/mgen.0.000294.
- Du Y, Sun F. 2022. HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome Biology* 23(1):63 DOI 10.1186/s13059-022-02626-w.
- Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461 DOI 10.1093/bioinformatics/btq461.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133–138 DOI 10.1126/science.1162986.
- Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32(19):3047–3048 DOI 10.1093/bioinformatics/btw354.
- Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, Furlanello C. 2018. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* 19(S2):49 DOI 10.1186/s12859-018-2033-5.
- Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S. 2020. LongQC: a quality control tool for third generation sequencing long read data. *G3 Genes|Genomes|Genetics* 10(4):1193–1196 DOI 10.1534/g3.119.400864.
- Gou W, Ling C, He Y, Jiang Z, Fu Y, Xu F, Miao Z, Sun T, Lin J, Zhu H, Zhou H, Chen Y, Zheng J-S. 2020. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 44(2):358–366 DOI 10.2337/dc20-1536.

- Gupta S, Kumar D, Negi R, Singh R, Ansari MJ, Kashyap S. 2024. *Genomic intelligence: metagenomics and artificial intelligence*. London: CRC Press. Available at <https://doi.org/10.1201/9781003570233>.
- Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* 68(4):669–685 DOI 10.1128/mmbr.68.4.669-685.2004.
- Hera MR, Liu S, Wei W, Rodriguez JS, Ma C, Koslicki D. 2024. Metagenomic functional profiling: to sketch or not to sketch? *Bioinformatics* 40:ii165–73 DOI 10.1093/bioinformatics/btae397.
- Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S. 2022. Machine learning and deep learning applications in microbiome research. *ISME Communications* 2(1):1–7 DOI 10.1038/s43705-022-00182-9.
- Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: an overview. *Human Immunology* 82(11):801–811 DOI 10.1016/j.humimm.2021.02.012.
- Huang N, Nie F, Ni P, Luo F, Wang J. 2022. SACall: a neural network basecaller for oxford nanopore sequencing data based on self-attention mechanism. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19(1):614–623 DOI 10.1109/TCBB.2020.3039244.
- Hugenholtz P, Tyson GW. 2008. Metagenomics. *Nature* 455(7212):481–483 DOI 10.1038/455481a.
- Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Kinder Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izaard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O’Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–214 DOI 10.1038/nature11234.

- Jain M, Olsen HE, Paten B, Akeson M. 2016. The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* 17(1):239 DOI 10.1186/s13059-016-1103-0.
- Jeon SA, Park JL, Park S-J, Kim JH, Goh S-H, Han J-Y, Kim S-Y. 2021. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes & Genomics* 43(7):713–724 DOI 10.1007/s13258-021-01096-x.
- Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28(1):27–30 DOI 10.1093/nar/28.1.27.
- Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7(11):e7359 DOI 10.7717/peerj.7359.
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* 41(1):e1 DOI 10.1093/nar/gks808.
- Kumar KR, Cowley MJ, Davis RL. 2024. Next-generation sequencing and emerging technologies*. *Seminars in Thrombosis and Hemostasis* 50(07):1026–1038 DOI 10.1055/s-0044-1786397.
- Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepille DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 31(9):814–821 DOI 10.1038/nbt.2676.
- Lassmann T. 2023. SAMStat 2: quality control for next generation sequencing data. *Bioinformatics* 39(1):btad019 DOI 10.1093/bioinformatics/btad019.
- Lecun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324 DOI 10.1109/5.726791.
- Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Research* 39:D19–D21 DOI 10.1093/nar/gkq1019.
- Levy Karin E, Steinegger M. 2025. Cutting-edge deep-learning based tools for metagenomic research. *National Science Review* 12(6):nwaf056 DOI 10.1093/nsr/nwaf056.
- Li J, Jew B, Zhan L, Hwang S, Coppola G, Freimer NB, Sul JH. 2019. ForestQC: quality control on genetic variants from next-generation sequencing data using random forest. *PLOS Computational Biology* 15(12):e1007556 DOI 10.1371/journal.pcbi.1007556.
- Lo C-C, Chain PSG. 2014. Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics* 15(1):366 DOI 10.1186/s12859-014-0366-2.
- Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* 21(10):597–614 DOI 10.1038/s41576-020-0236-x.
- Lundberg S, Lee S-I. 2017. A unified approach to interpreting model predictions. Available at <https://doi.org/10.48550/arXiv.1705.07874>.
- Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovic V, Aasmets O, Berland M, Gruca A, Hasic J, Hron K, Klammsteiner T, Kolev M, Lahti L, Lopes MB, Moreno V, Naskinova I, Org E, Paciência I, Papoutsoglou G, Shigdel R, Stres B, Vilne B, Yousef M, Zdravevski E, Tsamardinos I, Carrillo de Santa Pau E, Claesson MJ, Moreno-Indias I. 2021. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology* 12:105 DOI 10.3389/fmicb.2021.634511.
- McCombie WR, McPherson JD, Mardis ER. 2019. Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine* 9(11):a036798 DOI 10.1101/cshperspect.a036798.

- Metwally AA, Yu PS, Reiman D, Dai Y, Finn PW, Perkins DL. 2019. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via Long Short-Term Memory networks. *PLOS Computational Biology* 15(2):e1006693 DOI 10.1371/journal.pcbi.1006693.
- Mubaraki FA. 2025. Extensive novel diversity and phenotypic associations in the dromedary camel microbiome are revealed through deep metagenomics and machine learning. *PLOS ONE* 20(7):e0328194 DOI 10.1371/journal.pone.0328194.
- Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. 2021. Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* 39(5):555–560 DOI 10.1038/s41587-020-00777-4.
- Odrzywolek K, Karwowska Z, Majta J, Byrski A, Milanowska-Zabel K, Kosciolk T. 2022. Deep embeddings to comprehend and visualize microbiome protein space. *Scientific Reports* 12(1):10332 DOI 10.1038/s41598-022-14055-7.
- Oxford Nanopore Technologies. 2025a. Dorado. Available at <https://github.com/nanoporetech/dorado> (accessed 7 February 2025).
- Oxford Nanopore Technologies. 2025b. Medaka. Available at <https://github.com/nanoporetech/medaka> (accessed 7 February 2025).
- Oxford Nanopore Technologies. 2023. Nanopore Community Meeting 23 Tech Update. Available at <https://nanoporetech.com/en/news/news-oxford-nanopore-announces-breakthrough-performance-simplex-single-molecule-accuracy> (accessed 7 February 2025).
- Oxford Nanopore Technologies. 2024. Transforming basecalling in genomic sequencing. Available at <https://nanoporetech.com/blog/transforming-basecalling-in-genomic-sequencing> (accessed 7 February 2025).
- Pan S, Zhao X-M, Coelho LP. 2023. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 39:i21-9 DOI 10.1093/bioinformatics/btad209.
- Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLOS ONE* 7(2):e30619 DOI 10.1371/journal.pone.0030619.
- Pearson K. 2010. LIII. On lines and planes of closest fit to systems of points in space 1901. Available at <https://doi.org/10.1080/14786440109462720>.
- Pope Q, Varma R, Tataru C, David M, Fern X. 2023. Learning a deep language model for microbiomes: the power of large scale unlabeled microbiome data 2024. Available at <https://doi.org/10.1101/2023.07.17.549267>.
- Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13(1):341 DOI 10.1186/1471-2164-13-341.
- Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35(9):833–844 DOI 10.1038/nbt.3935.
- Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. 2016. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications* 469(4):967–977 DOI 10.1016/j.bbrc.2015.12.083.
- Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK. 2023. Next-generation sequencing technology: current trends and advancements. *Biology* 12(7):997 DOI 10.3390/biology12070997.
- Shah N, Molloy EK, Pop M, Warnow T. 2021. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics* 37(13):1839–1845 DOI 10.1093/bioinformatics/btab023.

- Sharpston TJ. 2014. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science* 5(e1002358):2629 DOI 10.3389/fpls.2014.00209.
- Stamatakis A. 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313 DOI 10.1093/bioinformatics/btu033.
- Stanojević D, Lin D, Nurk S, de SPF, Šikić M. 2024. Telomere-to-telomere phased genome assembly using HERRO-corrected simplex nanopore reads 2024. Available at <https://doi.org/10.1101/2024.05.18.594796>.
- Tamayo M, Olivares M, Ruas-Madiedo P, Margolles A, Espín JC, Medina I, Moreno-Arribas MV, Canals S, Mirasso CR, Ortín S, Beltrán-Sánchez H, Palloni A, Tomás-Barberán FA, Sanz Y. 2024. How diet and lifestyle can fine-tune gut microbiomes for healthy aging. *Annual Review of Food Science and Technology* 15(1):283–305 DOI 10.1146/annurev-food-072023-034458.
- Thireou T, Reczko M. 2007. Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4(3):441–446 DOI 10.1109/tcbb.2007.1015.
- Thomas T, Gilbert J, Meyer F. 2012. Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2(1):3 DOI 10.1186/2042-5783-2-3.
- Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research* 47(21):10994–11006 DOI 10.1093/nar/gkz841.
- Trivedi UH, Cézard T, Bridgett S, Montazam A, Nichols J, Blaxter M, Gharbi K. 2014. Quality control of next-generation sequencing data without a reference. *Frontiers in Genetics* 5:403 DOI 10.3389/fgene.2014.00111.
- van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics* 30(9):418–426 DOI 10.1016/j.tig.2014.07.001.
- Walsh C, Stallard-Olivera E, Fierer N. 2023. Nine (not so simple) steps: a practical guide to using machine learning in microbial ecology. *mBio* 15(2):e02050-23 DOI 10.1128/mbio.02050-23.
- Wang T, Wang X-W, Lee-Sarwar KA, Litonjua AA, Weiss ST, Sun Y, Maslov S, Liu Y-Y. 2023. Predicting metabolomic profiles from microbial composition through neural ordinary differential equations. *Nature Machine Intelligence* 5(3):284–293 DOI 10.1038/s42256-023-00627-3.
- Warburton PE, Sebra RP. 2023. Long-read DNA sequencing: recent advances and remaining challenges. *Annual Review of Genomics and Human Genetics* 24(1):109–132 DOI 10.1146/annurev-genom-101722-103045.
- Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* 13(6):e1005595 DOI 10.1371/journal.pcbi.1005595.
- Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. *PLOS Computational Biology* 6(2):e1000667 DOI 10.1371/journal.pcbi.1000667.
- Wu J, Singleton SS, Bhuiyan U, Krammer L, Mazumder R. 2024. Multi-omics approaches to studying gastrointestinal microbiome in the context of precision medicine and machine learning. *Frontiers in Molecular Biosciences* 10:3264 DOI 10.3389/fmolb.2023.1337373.
- Xie Z, Canalda-Baltrons A, d’Enfert C, Manichanh C. 2023. Shotgun metagenomics reveals interkingdom association between intestinal bacteria and fungi involving competition for nutrients. *Microbiome* 11(1):275 DOI 10.1186/s40168-023-01693-w.
- Yeh Y-M, Lu Y-C. 2022. MSRCall: a multi-scale deep neural network to basecall Oxford Nanopore sequences. *Bioinformatics* 38(16):3877–3884 DOI 10.1093/bioinformatics/btac435.

- Zhang W, Fan X, Shi H, Li J, Zhang M, Zhao J, Su X. 2023.** Comprehensive assessment of 16S rRNA gene amplicon sequencing for microbiome profiling across multiple habitats. *Microbiology Spectrum* **11**(3):e00563-23 DOI [10.1128/spectrum.00563-23](https://doi.org/10.1128/spectrum.00563-23).
- Zhang H, Jain C, Aluru S. 2020.** A comprehensive evaluation of long read error correction methods. *BMC Genomics* **21**(S6):889 DOI [10.1186/s12864-020-07227-0](https://doi.org/10.1186/s12864-020-07227-0).
- Zhou Z, Wu W, Ho H, Wang J, Shi L, Davuluri RV, Wang Z, Liu H. 2025.** DNABERT-S: pioneering species differentiation with species-aware DNA embeddings. *Bioinformatics* **41**:i255-64 DOI [10.1093/bioinformatics/btaf188](https://doi.org/10.1093/bioinformatics/btaf188).
- Ziemann M, Poulain P, Bora A. 2023.** The five pillars of computational reproducibility: bioinformatics and beyond. *Briefings in Bioinformatics* **24**(6):bbad375 DOI [10.1093/bib/bbad375](https://doi.org/10.1093/bib/bbad375).