OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf176
https://doi.org/10.1093/bib/bbaf176
Review

Bioinformatic approaches to blood and tissue microbiome analyses: challenges and perspectives

Jammi Prasanthi Sirasani[1,‡], Cory Gardner[2,‡], Gihwan Jung[2], Hyunju Lee[3], Tae-Hyuk Ahn[1,2,*]

[1]Program of Bioinformatics and Computational Biology, Saint Louis University, St. Louis, MO, United States
[2]Department of Computer Science, Saint Louis University, St. Louis, MO, United States
[3]AI Graduate School, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

*Corresponding author. E-mail: taehyuk.ahn@slu.edu
‡Jammi Prasanthi Sirasani and Cory Gardner contributed equally to this work.

Abstract

Advances in next-generation sequencing have resulted in a growing understanding of the microbiome and its role in human health. Unlike traditional microbiome analysis, blood and tissue microbiome analyses focus on the detection and characterization of microbial DNA in blood and tissue, previously considered a sterile environment. In this review, we discuss the challenges and methodologies associated with analyzing these samples, particularly emphasizing blood and tissue microbiome research. Key preprocessing steps—including the removal of ribosomal RNA, host DNA, and other contaminants—are critical to reducing noise and accurately capturing microbial evidence. We also explore how taxonomic profiling tools, machine learning, and advanced normalization techniques address contamination and low microbial biomass, thereby improving reliability. While it offers the potential for identifying microbial involvement in systemic diseases previously undetectable by traditional methods, this methodology also carries risks and lacks universal acceptance due to concerns over reliability and interpretation errors. This paper critically reviews these factors, highlighting both the promise and pitfalls of using blood and tissue microbiome analyses as a tool for biomarker discovery.

Keywords: blood and tissue; microbiome analysis; metagenomics; contamination; preprocessing; taxonomic profiling

Introduction

Human microbial communities reside in specialized living tissues such as the skin and gut, which are in constant interaction with the human body and influence health [1, 2]. While these highly colonized microbes have been widely studied, the characterization of low-biomass sites—such as certain tissues—has proven more challenging [3–5]. Recent advances in next-generation sequencing and other omics technologies have increased interest and debate in the detection of microbial DNA in tissues and blood, which were traditionally considered sterile environments [6, 7]. Tissue-resident microbes have been observed in both tumor and non-tumor contexts, leading to large-scale investigations to distinguish true microbial signals from contaminants [8]. For instance, a pan-cancer study by Dohlman et al. examined the 'Cancer Microbiome Atlas' to distinguish tissue-resident microbiota from background noise and highlighted how rigorous contamination filtering is essential to draw accurate conclusions in low-biomass samples [9]. These insights into tissue-based contamination issues also benefit the study of the blood microbiome, which has low microbial biomass and is particularly vulnerable to external and internal contamination [10, 11].

Building on growing evidence for microbiomes in sites previously deemed sterile, there is an increasing interest in leveraging blood and tissue microbiome analyses for disease biomarker discovery. In oncology, Poore et al. identified unique microbial signatures in both blood and tissues across a variety of cancer types [12], but this study was later retracted [13] following concerns raised by Gihawi et al., who argued that batch correction and database contamination with host sequences may have artificially created the appearance of cancer type-specific microbiomes [14]. In addition, the reproducibility of these findings has been debated in subsequent re-analyses [14, 15]. Other studies reported that the microbial signal in the blood of healthy individuals is highly transient, complicating efforts to distinguish clinically meaningful patterns from background noise [16]. The transient nature of the blood microbiome complicates the distinction between healthy and disease states due to the risk of contamination from low-biomass amplification and false positives of polymerase chain reaction (PCR) [6]. Mislabeled sequences in databases present additional challenges, especially in blood and tissue metagenomic analyses with insufficient microbial count [14, 17, 18].

Given the diagnostic potential of 'off-target' microbiomes, there remains a pressing need for bioinformatic approaches that address the unique challenges raised by low-biomass samples. Blood, for instance, contains high levels of host DNA relative to microbial DNA, whereas tissue samples can exhibit localized microbial features formed by diverse cellular environments [6]. Both have a high potential for contamination and false positives, highlighting the importance of specialized

Received: August 30, 2024. **Revised:** March 5, 2025. **Accepted:** March 25, 2025

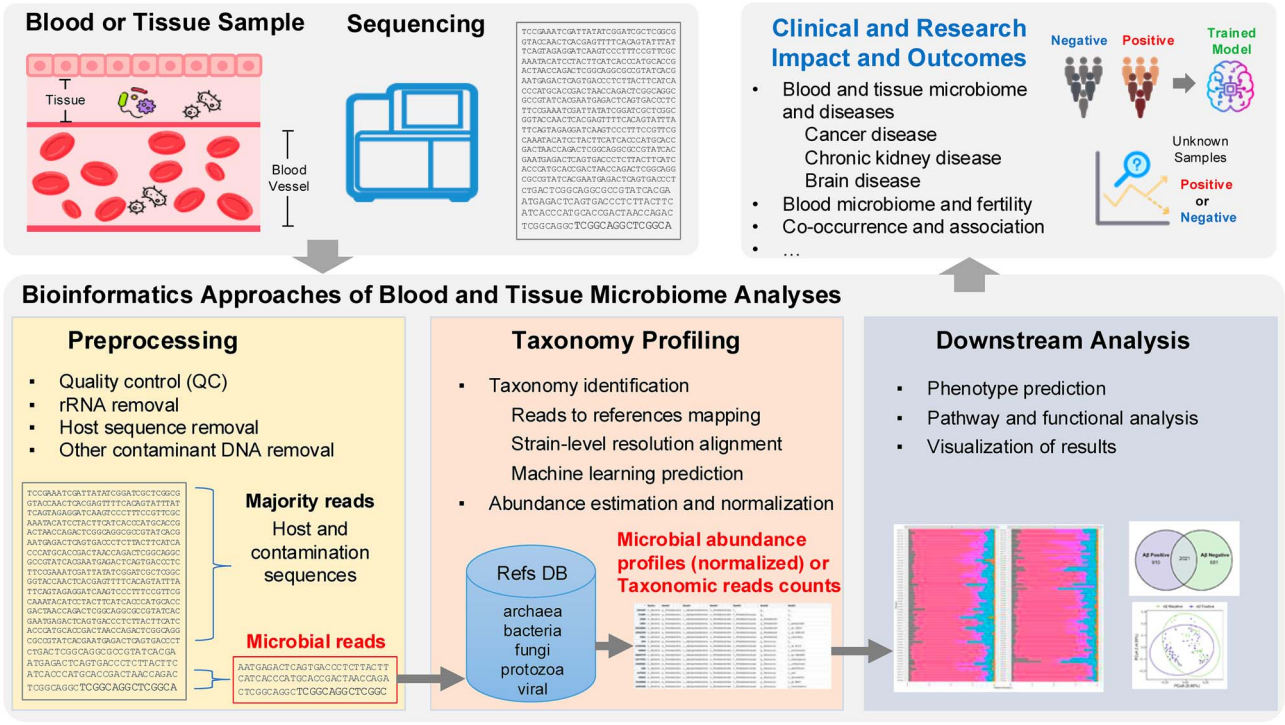

Figure 1. Overview of workflow for blood and tissue microbiome analyses. This figure outlines the methodology for studying the microbiome from blood and tissue samples, starting with sample preprocessing to remove host and contaminant sequences, followed by taxonomic profiling to determine the microbial composition, and concluding with downstream analysis for data interpretation.

computational pipelines for low-biomass data. In addition to improving sampling and sequencing strategies, comprehensive curation of reference databases—including the inclusion of newly discovered taxa—can significantly improve taxonomic resolution. Another promising research direction is the use of bioinformatic tools to validate polymorphism hypotheses from multiple perspectives based on large amounts of publicly available data. In this review, we examine key bioinformatic methodologies spanning preprocessing (host DNA depletion, contaminant filtering, and rRNA removal), taxonomic profiling, and downstream analyses (normalization, phenotype prediction, and data visualization).

By placing a particular emphasis on the blood and tissue microbiome studies, we highlight shared challenges that can enhance reliability and reproducibility in this emerging field. To ensure we present a comprehensive yet focused overview of relevant tools and references, we performed a structured literature search across databases such as PubMed, Web of Science, and Google Scholar, especially from 2019 to 2024. We included methods, pipelines, and studies that (i) explicitly address low-biomass genomic or transcriptomic workflows (e.g. blood and tissue), (ii) demonstrate widespread usage in recent microbiome research, or (iii) offer unique capabilities pertinent to contamination removal, taxonomic assignment, and normalization specific to low-biomass contexts. A schematic of the typical workflow for blood and tissue microbiome studies is shown in Fig. 1, highlighting the stages of sample preprocessing, taxonomic profiling, and downstream analysis. We generally excluded works that did not focus on blood or tissue samples, were largely duplicative of other references, or lacked sufficient evidence of validation.

General approaches to microbiome analysis

The term ‘microbiome’ has expanded over time to include not only microbes but also their collective genomes and interactions with the host and environment, but many studies still use ‘microbiome’ synonymously with the total microbial genetic content of a given location [1, 19]. In habitats with high biomass, such as the gut, researchers typically use well-established ‘meta-omics’ protocols such as metagenomics (microbial DNA) or metatranscriptomics (microbial RNA) to explore microbial composition and function [20, 21]. On the other hand, applying these same protocols to low-biomass niches (such as blood or certain tissues) requires greater caution due to the risk of contamination and the high ratio of host DNA to microbial DNA.

Clarifying ‘microbiome’ and ‘meta-omics’

While ‘metagenomics’ and ‘metatranscriptomics’ typically refer to the intentional sequencing of all microbial DNA or RNA in a community, microbial reads in blood or tissue studies can also occur incidentally as unmapped reads in host-centric whole genome or transcriptome datasets. This blurs the distinction between ‘meta-omics’ experiments and extended host sequencing approaches, creating semantic and technical challenges. In particular, the low microbial load of blood and tissue magnifies contamination issues and complicates the interpretation of single or rare microbial alignments.

Sequencing techniques for microbiome studies

The gut microbiome offers a prime example of a high-biomass environment in which microbes are abundant relative to host material [22]. Amplicon sequencing remains a foundational approach for preliminary profiling of microbial communities [23]. It is relatively inexpensive and relies on targeted PCR primers directed at conserved regions of the genome, typically the 16S or 18S ribosomal subunit. In this way, amplicon sequencing captures conserved taxonomic markers that researchers can use to estimate community composition at the genus or species level.

However, amplicon-based analyses are prone to PCR biases and only provide limited functional insights, as they sequence only a small portion of the genome of the organism.

Shotgun metagenomic sequencing overcomes many limitations of amplicon-based techniques by sequencing the total genetic material in a sample [24]. This provides a more complete picture of both taxonomic composition and functional gene content and can be used to detect viruses, fungi, and other non-bacterial microorganisms that the 16S/18S-based approaches miss. Shotgun metagenomic data can also be used for *de novo* assembly, followed by binning of contigs (and potentially reassembly or refinement) to generate metagenome-assembled genomes (MAGs), enabling downstream analyses of metabolic pathways [25, 26].

Metatranscriptomics extends whole genome sequencing (WGS) by focusing on the actively expressed genes, offering insight into the functional activity of microbial communities [27]. By sequencing total RNA, researchers can analyse gene expression, identify regulatory networks, and even gauge microbial responses to environmental or host factors. Although rRNA depletion is critical, gut samples typically contain enough microbial mRNA to yield robust data [28].

General bioinformatic analysis approaches

Numerous reviews and protocol resources describe best practices for metagenomic workflows, including pipeline selection, sample preparation guidelines, and software benchmarks [29–36]. Typically, researchers use rigorous preprocessing (quality control, read trimming, and host-read subtraction) to remove low-quality or irrelevant sequences and then proceed with taxonomic profiling including alignment-based references, alignment-free classification, marker-based approaches, or hybrid methods. If coverage allows, *de novo* assembly of reads into contigs can improve strain-level resolution and recovery of MAGs. Binning methods use sequence composition or differential abundance patterns to group contigs belonging to individual taxa. Following taxonomic assignment, functional and downstream analyses such as normalization, differential abundance testing, and metabolic pathway exploration help transform raw data into biologically meaningful insights. Finally, visualization and interpretation techniques help to communicate complex community structures more clearly. While these steps have been established primarily through the study of high-biomass samples, such as the gut microbiome, they provide the foundation for specialized adaptations needed for the unique challenges of studying the blood and tissue microbiome.

Extending to blood and tissue

Despite much lower microbial densities, blood and certain tissues can still contain detectable microbial DNA or RNA. Initiatives such as mBodyMap [37] show that common microbiome workflows such as quality control, host depletion, and taxonomic classification can be applied to these samples. However, the risk of contamination, extreme host-to-microbial DNA ratios, and poor annotation in databases require more stringent protocols than those typically used for gut or other high-biomass samples. In addition, assembly-based strategies are often impractical due to limited coverage, and read mapping of microbial references in blood and tissue data requires rigorous verification to rule out contamination.

In Sections Blood and tissue microbiome analysis and Challenges and discussion, we take a closer look at these special challenges of blood and tissue microbiome analyses, detailing preprocessing, taxonomic profiling, and downstream analysis for low-biomass situations.

Blood and tissue microbiome analyses

Understanding the potential presence and clinical significance of microbial DNA in blood and tissues is an area of growing research interest [13, 38, 39]. Samples deriving from bloodstream or tumor tissues, however, pose special challenges, such as high host DNA background, multiple contamination sources, and limited microbial biomass [40, 41]. This section discusses recommended best practices, methodological issues, and current controversies surrounding blood and tissue microbiome analyses.

In microbiome research, 'low-biomass' refers to samples with exceptionally low microbial DNA content relative to host DNA. This definition varies across tissue types, with blood typically exhibiting extremely low microbial presence, while other tissues may show more variability. The scarcity of microbial DNA in these samples presents significant analytical challenges, as even minor contamination can disproportionately influence results. Consequently, researchers must employ rigorous experimental protocols and specialized data analysis techniques to ensure the validity of their findings. This approach is crucial for accurately characterizing the microbiome in low-biomass environments and distinguishing true microbial signals from potential contaminants. The field continues to develop, with ongoing discussions about best practices and methodological refinements for studying blood and tissue microbiomes.

Bioinformatic preprocessing for low-biomass environments

Quality control, host subtraction, and contaminant removal

Low microbial abundance in blood and many tissue types means that host DNA can overwhelm microbial reads by orders of magnitude [42]. Therefore, host-subtraction pipelines (e.g. Kneaddata for short reads [(https://github.com/biobakery/kneaddata)] and Hostile for long reads [43]) are essential for accurate downstream analysis. Initial sequence quality assessment using FastQC [44] and MultiQC [45] is crucial for evaluating read quality metrics, duplication rates, and potential contamination before host depletion. Even very small amounts of extraneous bacterial DNA can be misidentified as low-abundance taxa. Methods such as Decontam [46], Recentrifuge [47], or Squeegee [48] apply statistical filters and negative controls to identify likely contaminants by examining read prevalence across blanks and experimental samples. Prevalence-based and batch-based filters also help flag microbes present exclusively in particular reagent batches. If negative controls are scarce, Squeegee offers a *de novo* approach to contaminant detection [48]. A detailed summary of prominent tools for host decontamination, contaminant filtering, and rRNA removal is found in Table 1.

rRNA removal in metatranscriptomics

In addition to bacterial rRNA, tissue- and blood-based metatranscriptomics must also contend with large amounts of host rRNA. Tools such as SortMeRNA [55] or Barrnap [53] target unwanted rRNA reads computationally, freeing resources to focus on mRNA that reflects true microbial activity. Because rRNA depletion can inadvertently remove non-rRNA microbial genes, verifying results with complementary data (i.e. total DNA) is advisable in borderline cases.

In summary, robust preprocessing including adapter trimming, host subtraction, and contaminant filtering lays a critical

Table 1. Preprocessing tools for low-biomass microbiome data

Tool [Citation]	Function/Algorithm	Pros	Cons
A. rRNA Read Identification and Removal			
RNAmmer [49]	HMM-based detection of full-length bacterial rRNA genes	Accurate for complete rRNAs	Not ideal for fragmented rRNA
Meta-RNA [50]	HMM approach specialized in identifying partial rRNA sequences	Good at handling rRNA fragments	Higher computational demands
rRNASelector [51]	HMM-driven detection of 5S, 23S, 26S rRNA genes	Tailored to longer rRNA subunits	Less commonly updated
Infernal 1.1 [52]	Covariance model (CM)-based RNA homology searches with fast filtering	High sensitivity for structured RNAs	Complex to configure for large datasets
Barrnap [53]	HMM-based tool for bacterial/archaeal rRNA gene prediction	Straightforward to run	Incompatible with latest HMMER updates
riboPicker [54]	Modified BWA-SW alignment to reference rRNA databases	Effective with partial rRNA reads	Can be slow for large datasets
SortMeRNA [55]	Approximate seed-based algorithm for rRNA filtering and OTU picking	Very fast filtering	May miss novel rRNA variants
rRNAFilter [56]	*k*-mer based expectation–maximization to distinguish rRNA from non-rRNA reads	No reference alignment needed	Parameter tuning can be challenging
RiboDetector [57]	BiLSTM DL model for identifying rRNA reads	High accuracy for partial/complete rRNA	Requires a suitable training set
B. Host Sequence Removal			
DeconSeq [58]	Mapping-based subtractive approach using BWA-SW	Simple workflow	No longer maintained; single-end only
MetaGeniE [59]	Multi-aligner strategy for filtering low-quality and human reads	Can detect multiple types of contaminants	Potentially high computational cost
CS-Score [60]	Heuristic pre-filter + directed mapping to partitioned human genome; cs-score to detect contaminants	Reduces false positives	Less common; depends on curated partitioned references
GenCoF [61]	Bowtie2-based multi-threaded host contamination filter	Fast for short Illumina reads	Not for long-read support
HoCoRT [62]	Short/long-read host contaminant removal via alignment + parameter optimization	Flexible read-length support	Relatively new; needs more benchmarking
Hostile [43]	Indexed approach for removing human reads while retaining microbes	Handles both short and long reads	Limited public benchmarks
Kneaddata (GitHubLink)	Bowtie2-based QC + decontamination pipeline with custom reference support	User-friendly; multi-threading support	Heavy computational time (dependent by DB)
HumanMycobiomeScan [63]	Identifies fungal reads by mapping to curated fungal databases	Specialized tool for mycobiome research	Limited to fungal references
ViromeScan [64] / ReadItAndKeep [65]	Focus on viral reads by aligning to known viral genomes	Targeted detection of viral communities	May miss novel/rare viruses
C. External and Internal Contaminant Removal			
Recentrifuge [47]	Uses negative controls + cross-sample comparison to subtract cross-contamination	Good for clinical or multiple-sample scenarios	Requires high-quality metadata/controls
Decontam [46]	Statistical classification of contaminants in R, based on frequency/prevalence models	Widely adopted in low-biomass studies	Focused mainly on 16S/amplicon data
MicroBIEM [66]	Negative-control ratio filter for suspected contaminants	Straightforward approach	Strongly depends on well-designed controls
Squeegee [48]	*De novo* approach for low-biomass contamination removal without negative controls	Ideal when controls are unavailable	Still in early adoption phase
SourceTracker / Meta-SourceTracker [67, 68]	Bayesian source tracking to identify proportion of contamination from various environmental/lab sources	Flexible for multiple contamination sources	Requires comprehensive source database
microDecon [69]	Read-subtraction approach comparing samples to blank controls	Easy concept	Can over-filter legitimate rare taxa
Prevalence / Correlation / Batch / Read-Counts Filters [16]	Simple statistical filters based on distribution patterns and negative-control prevalence	Quick to integrate into pipelines	Risk of removing genuine low-abundance taxa if thresholds are strict

foundation for low-biomass microbiome studies. By ensuring that remaining reads primarily reflect true microbial signals, researchers can proceed with greater confidence to taxonomic profiling.

Taxonomic profiling

In low-biomass samples, capturing enough microbial reads to confidently assign taxonomy is challenging. For 16S rRNA gene sequencing, commonly used tools such as QIIME2 [70] for taxonomic classification and DADA2 [71] for quality filtering, denoizing, and generating amplicon sequence variants can be employed. For WGS and RNA-seq samples, taxonomic profiling tools having large reference genome database and offering custom database options—such as Kraken 2 [72]—are particularly advantageous. These tools are better suited for detecting low-abundance microbial evidence compared to marker-based taxonomic profilers such as MetaPhlAn 4 [73] with limited marker-based databases. However, it is well recognized that such *k*-mer-based tools can overclassify spurious taxa in low-biomass samples [14]. Alignment-based strategies such as the BLAST-like tools DIAMOND [74], GATK PathSeq [75], and commercial DNAMAN software (https://www.lynnon.com/) can be used to reduce false positives resulting from a read sharing a *k*-mer with multiple taxa. Some researchers recommend combining alignment-free approaches with alignment-based approaches as "validation checks," and those approaches can improve the confidence of taxonomic assignments in low-biomass contexts [76, 77].

De novo assembly (SPAdes [78] or MEGAHIT [79]) followed by binning (MetaBAT [80] and CONCOCT [81]) can identify novel strains or species that are missing in reference databases, but low coverage in blood or tumor tissues often yields highly fragmented assemblies and makes assembly impractical for such contexts. Hybrid approaches—which map contigs to partially complete reference genomes—can help overcome this limitation while enabling identification of novel sequences.

For strain-level resolution, which is crucial for tracking disease outbreaks and antibiotic resistance mechanisms, several tools have been developed. Popular approaches include Pathoscope 2.0 [82], which provides comprehensive strain identification using alignment scores, SIGMA [83], which employs maximum likelihood estimation for accurate quantification, and StrainPhlAn 4 [84], which uses unique clade-specific marker genes. These tools enable researchers to distinguish subtle genetic differences within species, though their effectiveness in low-biomass samples requires careful validation.

Machine learning (ML) and deep learning (DL) are also applied for taxonomic profiling. DeepMicrobes [85] employs *k*-mer embedding and biLSTM networks for taxonomic classification of short metagenomic sequencing reads. BERTax [86] employs a deep neural network based on natural language processing to accurately classify DNA sequences at the superkingdom and phylum level without any known database representation. MT-MAG [87] is an ML-based tool designed for the taxonomic assignment of both complete and partial MAGs based on *k*-mer frequencies. CHEER [88] employs hierarchical convolutional neural networks (CNNs) and *k*-mer embedding for the taxonomic classification of viral genomes, especially in handling new species. DL-TODA [89] is another DL model employing CNNs that can support large data volumes. These approaches can be especially valuable in low-biomass settings, where traditional alignment-based methods may struggle with ambiguous signals. However, false positives are a major concern, and careful validation is still essential in low-biomass contexts. A detailed list of commonly used taxonomic profilers—including general-purpose, strain-level, and ML- and DL-based approaches—appears in Table 2.

To summarize, accurate taxonomic profiling in low-biomass contexts hinges on well-curated databases, careful choice of alignment-based or alignment-free, and stringent validation steps. Once microbial taxa are reliably identified, researchers can leverage downstream analyses such as functional characterization and phenotype prediction to unravel the clinical and biological implications of these taxa. The next section details these downstream steps, highlighting how integrated approaches further enhance the interpretation of blood and tissue microbiome data.

Downstream analysis

Following preprocessing and classification, downstream analysis seeks to extract clinically or biologically relevant signals from the data. This section highlights three key downstream applications that are especially relevant in blood and tissue contexts, where contamination and low coverage can confound standard approaches: phenotype prediction, functional/pathway analysis, and visualization.

Phenotype prediction

ML and DL have emerged as powerful tools for predicting disease states (e.g. cancer subtypes) based on microbiome composition. In blood and tissue microbiome studies, low microbial load and risk of contamination can severely affect predictive models if not carefully addressed. ML models should incorporate robust cross-validation (e.g. repeated stratified *k*-fold) and external validation cohorts. Without these, spurious patterns arising from batch effects or rare contaminants may falsely appear as significant biomarkers. After model training, identifying the highest-ranked taxa or features can reveal whether they stem from legitimate biological signals or artificially normalized read counts (as with the Poore et al. example). MetAML leverages multiple ML models to associate taxonomic profiles with specific phenotypes [104]. DeepMicro—a well-cited deep representation learning framework—effectively converts high-dimensional microbiome profiles into low-dimensional representations [105]. However, it often loses important features during this transformation, making it less suitable for low-biomass microbiome analysis. Tools such as SHAP [106] or LIME [107] facilitate interpretability, allowing researchers to confirm that model outputs align with biologically meaningful insights. Finally, comparing predictions made on raw data versus normalized data is recommended; major discrepancies could signal normalization-induced artifacts or other confounding factors.

Functional annotation and pathway analysis

Where coverage allows, functional analyses can shed light on microbial metabolic potential or gene expression. Aligning reads against protein databases (e.g. with DIAMOND [74]) or reconstructing functional profiles from the largest set of reference sequences (e.g. HUMAnN 3 [108]) helps identify putative metabolic pathways. The reliability of these predictions improves when supported by complementary transcriptomic data, ensuring that predicted genes are actively expressed. In disease contexts (e.g. comparing tumor versus adjacent normal tissue), researchers often apply specialized tools (e.g. DESeq2 [109] adapted for microbiome or ANCOM-BC [110]) to identify significantly enriched taxa or pathways. Caution is critical in low-biomass settings due to potential zero-inflation and contamination outliers.

Table 2. Taxonomic profiling tools for low-biomass microbiome data

Tool [Citation]	Function/Algorithm	Pros	Cons
General-Purpose Taxonomic Profiling			
Kraken2 [72]	Exact k-mer matching for metagenomic classification	Very fast classification	May overestimate taxa if reference is incomplete
PathSeq	Nucleotide-based alignment approach integrated with GATK for host read depletion and microbial detection	Flexible reference setups; strong integration with GATK tools	High computational overhead for large datasets; requires thorough reference curation
KrakenUniq [90]	Extension of Kraken with unique k-mer counting for improved abundance quantification	Reduces false positives in large data sets	Requires more memory than Kraken2 due to unique k-mer tracking
DIAMOND [74]	Protein and translated DNA alignment (blast-like)	Very high speed for large metagenomic databases	Limited to protein-level classification; no direct rRNA support
Kaiju [91]	Protein-level classification using a greedy search approach	Identifies organisms lacking annotated genomes	Might miss organisms with low-protein reference coverage
Mmseqs2 [92]	Large-scale protein search and clustering for taxonomic profiling	Highly scalable for big metagenomic data	Requires parameter tuning for sensitivity versus speed
mOTUs2 [93]	Marker gene-based profiling of microbial communities (metagenomic OTUs)	Good for species-level profiling of uncultured taxa	Focuses mainly on known marker genes; may miss rare lineages
Strain-Level Taxonomic Profiling			
PathoScope 2.0 [94]	Alignment-based framework for strain identification (scores alignment length + quality)	Enhanced precision for strain differentiation	Multiple components can be complex to set up
Sigma [95]	Nucleotide-level alignment + maximum-likelihood estimation for abundance quantification	Accurate genome abundance estimates	Less widely used; depends on robust reference genomes
StrainSeeker [96]	k-mer comparison to user-defined guide trees for rapid strain classification	Fast strain-level identification	Requires curated guide trees for optimal results
StrainEst [82]	Reference-based SNV profiling for multiple bacterial strains	Useful for focusing on a specific species of interest	Not a global classifier for all potential species
StrainGE [83]	Variant detection and strain analysis at low coverage (≥0.5x)	Works with minimal coverage	May miss ultra-rare strains with very low coverage
MetaPhlAn 4 [73]	Marker gene-based approach extended to strain-level profiling via StrainPhlAn	Widely validated pipeline	Primarily focuses on known clade-specific markers
CAMMiQ [97]	Variable-length substring analysis for single-cell or metagenomic strain identification	Flexible to different read lengths	Relatively new; performance under broad conditions less explored
StrainScan [98]	Tree-based k-mer indexing structure for precise strain-level analysis	Optimized for large-scale data sets	Requires significant computational memory for large references
StrainIQ [99]	n-gram-based algorithm for strain detection and abundance estimation	Effective at capturing subtle strain variants	Still emerging; not as extensively benchmarked
Strainberry [84]	Long-read metagenome strain separation and phasing	Leverages long-read advantages in resolving strains	Potentially high error rates if raw reads are noisy
Strainy [100]	Long-read assembly and phasing of multiple strains (Nanopore/HiFi)	Helpful for complex communities	Requires high-quality long reads
iGDA [101]	Detects and phases minor variants/strains from long-read data	Good for mixed populations with diverse haplotypes	Computationally intensive for large data sets
ML and DL in Taxonomic Classification			
DeepMicrobes [85]	DL (biLSTM) with k-mer embedding for short-read classification	High accuracy on partial/longer reads	Requires extensive training data
BERTax [86]	NLP-based approach (deep neural network) to classify DNA sequences, including novel species	Good at generalizing to unseen organisms	High computational cost; specialized hardware recommended
MT-MAG [87]	ML for partial or complete MAG taxonomy	Flexible classification for partial MAGs	Limited testing in extremely diverse microbiomes
CHEER [88]	Hierarchical CNN + k-mer embedding for viral classification	Efficient for new viral species	Focused on viral data; limited use for bacterial classification
DL-TODA [89]	Convolutional neural network for species-level classification in large omics data	Scales to massive data sets	Still emerging; fewer real-world benchmarks
MetaMaps [102]	Approximate mapping + scoring for long-read metagenomics with strain-level resolution	Handles high-error-rate long reads well	Reference-dependent; can be slow for very large databases
CDKAM [103]	Discriminative k-mer method for classifying third-generation sequencing reads	Robust to high-error long reads	Relatively new; not yet widely benchmarked

Visualization and communication of results

Pie charts, bar plots, or Krona [111] plots show the proportional abundance of microbial taxa and provide a clear way to visualize sample taxa composition. If used, color-coding can highlight possible contaminants flagged by negative controls. Tools such as iTOL [112] enable interactive display of phylogenetic trees, which is particularly helpful in identifying unknown or novel branches discovered via assembly-based approaches. For functional data, MetScape [113] within Cytoscape [114] can illustrate biochemical networks linking microbial genes, metabolites, and host factors, aiding comprehension of potential microbe-host interactions. Table 3 provides an overview of key tools for phenotype prediction, pathway analysis, and data visualization, with special attention given to low-biomass metagenomic data.

By employing phenotype prediction, functional pathway analysis, and clear visualization techniques, researchers can transform raw microbiome profiles into clinically or biologically meaningful insights. However, for samples with low biomass, more careful attention is needed to exclude artifacts and contamination. The following discussion addresses these broader challenges and emerging solutions.

Challenges and discussion

Blood and tissue microbiome analyses are increasingly an important topic in biomedical research, driven by the possibility of identifying novel disease biomarkers and therapeutic targets in what were traditionally believed to be 'sterile' environments. Recent studies have shown that microbial communities may exist in various tissue types and blood circulation, with potential impacts on both health and disease states [9, 38, 134, 135]. While the gut microbiome's role in human health is well-established, the presence and potential function of microbes in blood and tissues represents a new frontier in microbiome science. These investigations have revealed possible connections between tissue-specific microbial signatures and various pathological conditions, including cancer [9, 38], Parkinson's disease [135], and even autism [134].

However, the field remains contentious, with some large-scale studies questioning the existence of a consistent blood microbiome in healthy individuals [16], while others demonstrate tissue-specific microbial patterns in disease states [9]. This complexity is compounded by significant technical challenges in sample collection, processing, and analysis, particularly given the extremely low microbial biomass in these environments [41]. In this section, we discuss the key challenges and recent developments in blood and tissue microbiome analyses, including persistent technical hurdles in low-biomass contexts, taxonomic classification complexities, abundance estimation difficulties, the promise of multi-omic approaches, ongoing debates about core microbiomes, and the implications for clinical applications.

Persistent challenges in low-biomass contexts

A defining characteristic of blood and tissue samples is the very low ratio of microbial to host DNA. Even minimal contamination can hide genuine microbial signals and lead to false positives. Rigorous use of negative controls (e.g. reagent-only blanks) and robust host-subtraction pipelines (e.g. Kneaddata (https://github.com/biobakery/kneaddata) and Hostile [43]) can help validate findings and mitigate such risks. At the same time, high host DNA content complicates *de novo* assembly, which limits the ability to achieve strain-level resolution or to reconstruct MAGs. Alignment-based and reference-guided methods often prove more feasible in blood and tissue contexts, but they rely on the completeness and accuracy of reference databases, which often contain mislabeled or contaminated entries [14].

Taxonomic classification and database choice

Accurately defining the microbial signal in low-biomass contexts is complicated by the dual risks of incomplete or contaminated reference databases and insufficiently comprehensive host references. Although rapid classifiers such as Kraken2 [72] and other alignment-based or marker-based tools have advanced the field, recent evidence from Gihawi et al. [14] shows how omitting the human genome and including draft microbial assemblies can drastically inflate microbial read counts. Specifically, if the database lacks the human reference, any residual host reads that remain after imperfect human filtering can be erroneously assigned to microbial taxa. This phenomenon is amplified when draft genomes in public repositories contain mislabeled human sequences, a well-documented issue that has introduced spurious sequences into thousands of putative bacterial entries [14]. The net effect is that human reads appear to map exclusively to the microbial database, thus creating false positives and artificially high microbial abundances.

Species abundance estimation and sample normalization

Accurate quantification of microbial species in low-biomass samples is another challenge. Tools such as MEGAN [136] and Bracken [137] estimate abundance across taxonomic levels, while marker gene-based strategies (e.g. MetaPhlAn 4 [73]) focus on conserved regions for more reliable detection. However, these methods can yield spurious results if the underlying data are heavily zero-inflated or if key marker genes are missing [138]. Statistical packages such as ANCOM-BC [110] and ALDEx2 [139] were developed to control for compositional biases and variance in metagenomic data, but their success in low-biomass contexts demands careful thresholds for detection limits and false positives.

In addition, normalization practices and batch corrections further complicate analysis. As Gihawi et al. [14] illustrate, a misguided normalization approach may introduce tumor-specific or sample-specific signals directly into abundance matrices, which makes it trivial for ML algorithms to separate clinical categories, even when the taxa have no true reads in the original dataset. While these artifacts can appear in any low-biomass study, blood and tissue analyses are particularly susceptible, given the high ratio of host DNA to microbial DNA and the reliance on small statistical signals to detect real microbes. Researchers can help mitigate these pitfalls by including complete host references in taxonomic classification, excluding poorly curated or draft microbial assemblies known to have contamination, and carefully checking normalization pipelines to ensure they do not introduce artificial between-sample variation. In addition, best-practice workflows typically include multiple negative controls and re-check unexpected taxa against alternative databases to confirm whether a putative organism is truly present in the sample.

Integrating multi-omics and advanced machine learning and deep learning

Multi-omic integration is revolutionizing our understanding of host–microbiome interactions in health and disease. A recent study of early colorectal cancer used ML to combine genomic, transcriptomic, and microbiome data to study key drivers of tumor formation and progression [140]. For example, the HoloOmic approach [141] provides a comprehensive framework

Table 3. Downstream analysis tools for low-biomass microbiome data

Tool [Citation]	Function/Algorithm	Pros	Cons
Phenotype Prediction			
DeepMicro [105]	Deep representation learning framework for microbiome data	Well-cited and widely recognized framework	May lose important features during dimensionality reduction
MegaR [115]	ML-based sample classification from 16S/WGS data	Supports multiple ML models	Primarily tested on bacterial datasets
MegaD [116]	DL for predicting disease status from metagenomic profiles	Accurate on diverse microbial communities	Resource-intensive; larger training sets needed
MetaAnalyst [117]	Biomarker detection and phenotype prediction from metagenomic data	Focused on quick biomarker identification	May require extensive parameter tuning for complex data
MEGMA [118]	Converts metagenomic data into 2D 'microbiome prints' for DL with CNNs	Captures complex microbe interactions	Relies on consistent data transformation procedures
Meta-Singer [119]	Rank aggregation of ML-identified taxa to create a single prioritized biomarker list	Integrates multiple model outputs	Can be sensitive to incorrect model rankings
Ph-CNN [120]	Convolutional neural network approach to phenotype prediction based on phylogenetic features	Accounts for evolutionary relationships	Focused more on bacterial profiles; less tested on viruses
Pathway and Functional Analysis			
PICRUSt/PICRUSt2 [121, 122]	Predicts functional capacities from 16S data (marker-based metagenome prediction)	Easy integration with 16S workflows	Limited to bacterial/archaeal pathways; no direct read alignment
PanFP [123]	Pangenome-based functional profiling for microbial communities	Good for exploring functional diversity	Focused on known reference pangenomes
paprica [124]	Pathway prediction using phylogenetic placement	Useful for ecological studies	Relies on robust phylogenetic trees
Tax4Fun/Tax4Fun2 [125, 126]	Maps 16S rRNA data to reference genomes for functional inference	Quick pipeline for functional predictions	Dependent on accurate 16S-to-genome mappings
Piphillin [127]	Direct inference of functional profiles from 16S data	Simple cloud/web-based interface	Accuracy drops if references are distant from sample species
MicFunPred [128]	Predicts functional traits from 16S sequences, extends coverage to eukaryotes/viruses	Broader taxonomic scope	Still relatively new; fewer validations
DIAMOND [74]	Fast aligner for functional annotation against protein references	Extremely fast compared to BLAST	Protein level only; not suited for rRNA-based functional calls
COGNIZER [129]	Functional annotation using the COG (Clusters of Orthologous Groups) database	Integrates well with metatranscriptomic data	Heavily reliant on COG completeness for certain microbial groups
GHOSTX [130]	Rapid homology search tool for metagenomic sequence annotation	Faster alternative to BLASTX	May be less sensitive for highly divergent sequences
DeepFRI [131]	DL-based functional annotation of protein sequences	Can identify novel protein functions	Requires GPU resources; large training set needed
Visualization of Results			
MetScape [113]	Integrates metabolomic data with gene/pathway information for network views	Good for system-level metabolic insights	Needs frequent updates to match new metabolic databases
Krona [111]	Interactive radial tree visualizations of taxonomic/abundance data	Easy web integration; highly intuitive display	Limited advanced analysis features
BacMap [132]	Electronic atlas of bacterial genomes with interactive chromosome maps	Detailed genome visualization	Primarily focuses on prokaryotes; less coverage of eukaryotes/viruses
IMG/M [133]	Comprehensive platform for analysing and comparing microbial genomic datasets	Offers numerous built-in analytical pipelines	Requires data submission or setup in the environment
iTOL [112]	Web-based phylogenetic tree visualization and annotation	Highly interactive and customizable	Handles primarily tree-based data; limited non-phylogenetic features

Protocol	Challenges	Mitigation Strategies
Quality Control	• High adapter proportion in low-biomass samples • PCR biases amplifying fewer microbial reads	• Advanced QC algorithms (FastQC, MultiQC) • Optimization of amplification protocols
rRNA removal	• rRNA overabundance masks mRNA signals • PCR amplification artifacts	• Targeted rRNA depletion kits (SortMeRNA, Barrnap) • Verify with total DNA data in borderline cases
Host DNA removal	• High ratio of host to microbial DNA • Mislabeled sequences in reference databases	• Stringent mapping to human reference (Kneaddata) • Multiple passes of alignment
Contaminant removal	• External vs. internal contaminants difficult to differentiate • Reagent blanks, batch effects	• Incorporate negative controls • Use specialized tools (Decontam, Squeegee)
Taxonomy identification	• High contamination produces false positive • Incomplete or absent reference genomes	• Curation of databases (Kraken 2, PathSeq) • Combine alignment-based and alignment-free strategies
Assembly and binning	• Limited coverage in low-biomass samples • Host DNA overshadowing microbial reads	• Deep sequencing or selective capture of microbial DNA • Develop alternate strategies employing alignment-based or ML/DL models
Normalization	• Zero-inflation & batch-to-batch variability • Potential for introducing artificial signals	• Compare raw vs. normalized data • Employ validated pipelines (ANCOM-BC, ALDEx2)
Phenotype prediction	• Spurious associations from contaminants & batch effects • Confounding due to host factors (age, genotype)	• Include diverse datasets to train ML models • Use feature selection techniques to identify informative genetic markers (DeepMicro)
Pathway/functional analysis	• Contamination & normalization errors can mislead functional signals • Gaps in reference database	• Use comprehensive and curated databases (DIAMOND, ANCOM) • Employ statistical models that can handle sparse data

Figure 2. Overview of recommended protocols, major challenges, and possible mitigation strategies for blood and tissue microbiome analyses. Each step, from quality control and host DNA removal to normalization and phenotype prediction, presents unique difficulties in low-biomass environments. However, targeted solutions (e.g. optimized host subtraction, rigorous negative controls, and specialized bioinformatic tools) can improve confidence in microbial profiles and subsequent downstream insights.

for analyzing multi-omic data from both host and microbial domains, enabling deeper insights into their bidirectional interactions. In low-biomass settings, where microbial signals are often confounded by contamination or host-dominant features, combining metagenomics and metatranscriptomics with host-centric data (e.g. transcriptomics or proteomics) can enhance the detection of true microbial signatures. Effective integration requires careful study design and sample preparation, and strong quality control to ensure data reliability. By simultaneously examining microbial and host compartments, this approach can reveal metabolic and regulatory cross talk that is often missed in single omic analyses.

Nevertheless, the high dimensionality and potential zero-inflation of multi-omic data can raise significant computational challenges, particularly in low-biomass samples [142]. Advanced ML and DL techniques—such as autoencoders, variational inference, and graph-based fusion—address these challenges by extracting meaningful features and accommodating heterogeneous data [143]. These methods enhance biomarker discovery, classification accuracy, and understanding of dynamic host–microbe interactions. Rigorous controls and batch effect corrections are critical to ensure that findings reflect true biological signals rather than artifacts. Multi-omic data utilizing ML and DL can provide powerful directions for discovering clinically relevant biomarkers in blood and tissue microbiome studies, elucidating host–microbe mechanisms, and advancing disease diagnosis and treatment.

Debates on the 'Core' blood microbiome

One of the more contentious topics in the field is whether a stable 'core' community truly exists in the bloodstream. Studies such as Tan et al. and others argue that no consistent microbial signal emerges in healthy populations, suggesting that many reported 'blood microbes' could be contaminants or transient bacteria from the gut or skin [10, 16, 144]. By contrast, smaller investigations claim to detect recurring microbial signatures that may modulate immune responses [11]. Resolving this debate may require large-scale population studies incorporating systematic contamination controls, stratified sampling, and highly sensitive bioinformatic pipelines.

Clinical and translational implications

Given the contamination risks in low-biomass samples, researchers increasingly recommend a suite of measures to detect and control both cross-contamination and reagent contamination [16, 41]. Davis et al. [40] demonstrated that statistical approaches can identify and remove contaminant sequences, while Eisenhofer et al. [41] provided guidelines for handling contamination in low microbial biomass studies. The use of multiple negative controls (e.g. reagent blanks, mock extractions) improves the identification of reagent-derived taxa. Tools such as Decontam [46] and Squeegee [48] rely on these controls to statistically model background contaminants and reduce false-positive microbial calls.

Recent methodological advances have improved our ability to distinguish tissue-resident microbes from contaminants. Dohlman et al. [9] provided the 'Cancer Microbiome Atlas' approach, which uses comparative analysis across tissue types combined with contamination filtering. This method aids in discriminating between true tissue-resident microbiota and background noise in low-biomass samples through multi-step validation procedures and statistical controls.

Emerging technologies are also advancing our capabilities in low-biomass microbiome analysis. Long-read sequencing technologies continue to improve strain-level resolution, with both Oxford Nanopore [145] and PacBio HiFi [146] platforms achieving sufficient accuracy for reliable strain typing [100].

However, unaddressed issues remain: standard guidelines for how many negative controls to include (per batch or per sample set) are inconsistent across studies, and no universal threshold exists for removing contaminants without inadvertently discarding genuine low-abundance taxa [41]. Additionally, many alignment-based pipelines depend on well-annotated and contamination-free reference databases, yet mislabeled or incomplete entries in public repositories remain prevalent [17, 18]. Finally, the field lacks standardized workflows for sample processing, sequencing, and analysis, though initiatives are working to develop standard operating procedures [41].

Despite these challenges, blood and tissue metagenomic analyses could provide a powerful tool for clinical diagnostics and personalized medicine. For instance, the sensitive detection of microbial signatures in cancer patients may refine prognosis or guide targeted therapies [9, 38]. In other disease contexts, such as autoimmune disorders or chronic infections, understanding tissue-resident versus transient microbes could lead to new interventions. However, any clinical application must adopt standardized workflows, transparent reporting of negative controls, and robust normalization checks to instill confidence in the results.

As summarized in Fig. 2, a structured workflow with strict quality control, comprehensive host DNA removal, and careful contaminant filtering is essential for reliable detection of microbial signatures in low-biomass samples. By highlighting key pitfalls at each step, including background contamination, incomplete reference databases, and over-normalization, Figure 2 highlights both the complexity of blood and tissue microbiome analysis and the strategies needed to mitigate false signals.

Conclusion and future directions

Blood and tissue microbiome analyses hold great promise for identifying novel biomarkers, increasing our understanding of host–microbe interactions, and probing the roles of previously undetected microbial residents in systemic diseases. However, these studies are often constrained by low microbial biomass, which magnifies the impact of potential contaminants, incomplete reference databases, and normalization biases. Progress will be driven by large-scale population studies implementing standardized sampling protocols and comprehensive negative controls, as well as ongoing efforts to curate and improve reference genomes to minimize spurious host matches and mislabeled entries. Multi-omic approaches—which use complementary metatranscriptomic, metaproteomic, and metabolomic data—can further clarify whether detected microbes are metabolically active or merely present as cell-free DNA. Long-read sequencing technologies—such as Oxford Nanopore and PacBio HiFi—may help overcome assembly fragmentation and improve strain-level resolution. It is equally important to adopt interpretable ML methods to create high-accuracy classifiers that reflect real microbial variation rather than data processing artifacts. This refinement in best practices promises to determine, more decisively, whether the microbes detected in these nominally 'sterile' sites have true clinical significance or instead prove to be transient contaminants without functional impact.

Key Points

- Blood and tissue samples typically contain very low microbial biomass relative to host DNA and therefore require stringent decontamination and host-removal steps to prevent false positives. Even minimal reagent or laboratory contaminants can make it difficult to discover true microbial signals in low-biomass situations, highlighting the importance of negative controls and robust filtering tools.
- Mislabeled sequences in public reference genomes can artificially inflate microbial abundances. The entire human genome should be included in the classification pipeline, while careful attention should be given toward excluding poorly curated draft microbial assemblies.
- Normalization and batch corrections can inadvertently embed sample-specific artifacts in abundance data, potentially yielding deceptively high classification performance but little biological validity. Checking both raw and normalized data for consistency can reveal such artifacts.
- Machine learning and deep learning can detect and extract subtle microbial signals in blood and tissue samples, but robust cross-validation, external validation cohorts, and interpretability methods are essential to distinguish legitimate biomarkers from noise.
- Despite these challenges, identifying tissue-specific or blood-borne microbes holds promise for new diagnostic methods, improved prognostic markers, and targeted therapies but requires the use of validated workflows, contamination controls, and accurate database references.

Conflict of interest: None declared.

Funding

TA is supported by NSF-2430236.

References

1. Berg G, Rybakova D, Fischer D. *et al.* Microbiome definition re-visited: Old concepts and new challenges. *Microbiome* 2020;**8**:103. https://doi.org/10.1186/s40168-020-00875-0
2. Ogunrinola GA, Oyewale JO, Oshamika OO. *et al.* The human microbiome and its impacts on health. *Int J Microbiol* 2020;**2020**: 8045646–7. https://doi.org/10.1155/2020/8045646
3. Link CD. Is there a brain microbiome? *Neurosci Insights* 2021;**16**:26331055211018709. https://doi.org/10.1177/26331055211018709
4. Molina NM, Sola-Leyva A, Haahr T. *et al.* Analysing endometrial microbiome: Methodological considerations and recommendations for good practice. *Hum Reprod* 2021;**36**:859–79. https://doi.org/10.1093/humrep/deab009
5. Peric A, Weiss J, Vulliemoz N. *et al.* Bacterial colonization of the female upper genital tract. *Int J Mol Sci* 2019;**20**:3405. https://doi.org/10.3390/ijms20143405
6. Selway CA, Eisenhofer R, Weyrich LS. Microbiome applications for pathology: Challenges of low microbial biomass samples during diagnostic testing. *J Pathol Clin Res* 2020;**6**:97–106. https://doi.org/10.1002/cjp2.151
7. Wensel CR, Pluznick JL, Salzberg SL. *et al.* Next-generation sequencing: Insights to advance clinical investigations of the microbiome. *J Clin Invest* 2022;**132**:e154944. https://doi.org/10.1172/JCI154944
8. Yang J, Moon HE, Park HW. *et al.* Brain tumor diagnostic model and dietary effect based on extracellular vesicle microbiome data in serum. *Exp Mol Med* 2020;**52**:1602–13. https://doi.org/10.1038/s12276-020-00501-x
9. Dohlman AB, Arguijo Mendoza D, Ding S. *et al.* The cancer microbiome atlas: A pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* 2021;**29**:281–298 e285. https://doi.org/10.1016/j.chom.2020.12.001
10. Castillo DJ, Rifkin RF, Cowan DA. *et al.* The healthy human blood microbiome: Fact or fiction? *Front Cell Infect Microbiol* 2019;**9**:148. https://doi.org/10.3389/fcimb.2019.00148
11. Whittle E, Leonard MO, Harrison R. *et al.* Multi-method characterization of the human circulating microbiome. *Front Microbiol* 2018;**9**:3266. https://doi.org/10.3389/fmicb.2018.03266
12. Poore GD, Kopylova E, Zhu Q. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 2020;**579**:567–74. https://doi.org/10.1038/s41586-020-2095-1
13. Poore GD, Kopylova E, Zhu Q. *et al.* Retraction note: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 2024;**631**:694. https://doi.org/10.1038/s41586-024-07656-x
14. Gihawi A, Ge Y, Lu J. *et al.* Major data analysis errors invalidate cancer microbiome findings. *MBio* 2023;**14**:e0160723. https://doi.org/10.1128/mbio.01607-23
15. Sepich-Poore GD, McDonald D, Kopylova E. *et al.* Robustness of cancer microbiome signals over a broad range of methodological variation. *Oncogene* 2024;**43**:1127–48. https://doi.org/10.1038/s41388-024-02974-w
16. Tan CCS, Ko KKK, Chen H. *et al.* No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. *Nat Microbiol* 2023;**8**:973–85. https://doi.org/10.1038/s41564-023-01350-w
17. Breitwieser FP, Pertea M, Zimin AV. *et al.* Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* 2019;**29**:954–60. https://doi.org/10.1101/gr.245373.118
18. Steinegger M, Salzberg SL. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* 2020;**21**:115. https://doi.org/10.1186/s13059-020-02023-1
19. Marchesi JR, Ravel J. The vocabulary of microbiome research: A proposal. *Microbiome* 2015;**3**:31. https://doi.org/10.1186/s40168-015-0094-5
20. Dai X, Shen L. Advances and trends in omics technology development. *Front Med (Lausanne)* 2022;**9**:911861. https://doi.org/10.3389/fmed.2022.911861
21. Aguiar-Pulido V, Huang W, Suarez-Ulloa V. *et al.* Metagenomics, Metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evol Bioinform Online* 2016;**12**:5–16. https://doi.org/10.4137/EBO.S36436
22. Cani PD. Human gut microbiome: Hopes, threats and promises. *Gut* 2018;**67**:1716–25. https://doi.org/10.1136/gutjnl-2018-316723
23. Franzosa EA, Hsu T, Sirota-Madi A. *et al.* Sequencing and beyond: Integrating molecular 'omics' for microbial community profiling. *Nat Rev Microbiol* 2015;**13**:360–72. https://doi.org/10.1038/nrmicro3451
24. Ranjan R, Rani A, Metwally A. *et al.* Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* 2016;**469**: 967–77. https://doi.org/10.1016/j.bbrc.2015.12.083
25. Wang Z, Huang P, You R. *et al.* MetaBinner: A high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol* 2023;**24**:1. https://doi.org/10.1186/s13059-022-02832-6
26. Du Y, Sun F. HiCBin: Binning metagenomic contigs and recovering metagenome-assembled genomes using hi-C contact maps. *Genome Biol* 2022;**23**:63. https://doi.org/10.1186/s13059-022-02626-w
27. Franzosa EA, McIver LJ, Rahnavard G. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 2018;**15**:962–8. https://doi.org/10.1038/s41592-018-0176-y
28. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of Metatranscriptomics in microbiome research. *Bioinform Biol Insights* 2016;**10**:19–25. https://doi.org/10.4137/BBI.S34610
29. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2019;**20**:1125–36. https://doi.org/10.1093/bib/bbx120
30. Knight R, Vrbanac A, Taylor BC. *et al.* Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018;**16**:410–22. https://doi.org/10.1038/s41579-018-0029-9
31. McIntyre ABR, Ounit R, Afshinnekoo E. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* 2017;**18**:182. https://doi.org/10.1186/s13059-017-1299-7
32. Meyer F, Fritz A, Deng ZL. *et al.* Critical assessment of metagenome interpretation: The second round of challenges. *Nat Methods* 2022;**19**:429–40. https://doi.org/10.1038/s41592-022-01431-4
33. Quince C, Walker AW, Simpson JT. *et al.* Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**: 833–44. https://doi.org/10.1038/nbt.3935
34. Sczyrba A, Hofmann P, Belmann P. *et al.* Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods* 2017;**14**:1063–71. https://doi.org/10.1038/nmeth.4458

35. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* 2012;**2**:3. https://doi.org/10.1186/2042-5783-2-3
36. Yue Y, Huang H, Qi Z. *et al.* Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 2020;**21**:334. https://doi.org/10.1186/s12859-020-03667-3
37. Jin H, Hu G, Sun C. *et al.* mBodyMap: A curated database for microbes across human body and their associations with health and diseases. *Nucleic Acids Res* 2022;**50**:D808–16. https://doi.org/10.1093/nar/gkab973
38. Nejman D, Livyatan I, Fuks G. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 2020;**368**:973–80. https://doi.org/10.1126/science.aay9189
39. Robinson KM, Crabtree J, Mattick JS. *et al.* Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data. *Microbiome* 2017;**5**:9. https://doi.org/10.1186/s40168-016-0224-8
40. Davis NM, Proctor DM, Holmes SP. *et al.* Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 2018;**6**:226. https://doi.org/10.1186/s40168-018-0605-2
41. Eisenhofer R, Minich JJ, Marotz C. *et al.* Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends Microbiol* 2019;**27**:105–17. https://doi.org/10.1016/j.tim.2018.11.003
42. Zhang C, Cleveland K, Schnoll-Sussman F. *et al.* Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biol* 2015;**16**:265. https://doi.org/10.1186/s13059-015-0821-z
43. Constantinides B, Hunt M, Crook DW. Hostile: Accurate decontamination of microbial host sequences. *Bioinformatics* 2023;**39**:btad728. https://doi.org/10.1093/bioinformatics/btad728
44. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
45. Ewels P, Magnusson M, Lundin S. *et al.* MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8. https://doi.org/10.1093/bioinformatics/btw354
46. McArdle AJ, Kaforou M. Sensitivity of shotgun metagenomics to host DNA: Abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiol* 2020;**2**:acmi000104. https://doi.org/10.1099/acmi.0.000104
47. Marti JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput Biol* 2019;**15**:e1006967. https://doi.org/10.1371/journal.pcbi.1006967
48. Liu Y, Elworth RAL, Jochum MD. *et al.* De novo identification of microbial contaminants in low microbial biomass microbiomes with squeegee. *Nat Commun* 2022;**13**:6799. https://doi.org/10.1038/s41467-022-34409-z
49. Lagesen K, Hallin P, Rodland EA. *et al.* RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;**35**:3100–8. https://doi.org/10.1093/nar/gkm160
50. Huang Y, Gilna P, Li W. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 2009;**25**:1338–40. https://doi.org/10.1093/bioinformatics/btp161
51. Lee JH, Yi H, Chun J. rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J Microbiol* 2011;**49**:689–91. https://doi.org/10.1007/s12275-011-1213-z
52. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5. https://doi.org/10.1093/bioinformatics/btt509
53. Seemann T. (2014). barrnap: Bacterial Ribosomal RNA Predictor. [Online]. Available online at: https://github.com/tseemann/barrnap.
54. Schmieder R, Lim YW, Edwards R. Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 2012;**28**:433–5. https://doi.org/10.1093/bioinformatics/btr669
55. Kopylova E, Noe L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012;**28**:3211–7. https://doi.org/10.1093/bioinformatics/bts611
56. Wang Y, Hu H, Li X. rRNAFilter: A fast approach for ribosomal RNA read removal without a reference database. *J Comput Biol* 2017;**24**:368–75. https://doi.org/10.1089/cmb.2016.0113
57. Deng ZL, Munch PC, Mreches R. *et al.* Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res* 2022;**50**:e60. https://doi.org/10.1093/nar/gkac112
58. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One* 2011;**6**:e17288. https://doi.org/10.1371/journal.pone.0017288
59. Rawat A, Engelthaler DM, Driebe EM. *et al.* MetaGeniE: Characterizing human clinical samples using deep metagenomic sequencing. *PloS One* 2014;**9**:e110915. https://doi.org/10.1371/journal.pone.0110915
60. Haque MM, Bose T, Dutta A. *et al.* CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. *Genomics* 2015;**106**:116–21. https://doi.org/10.1016/j.ygeno.2015.04.005
61. Czajkowski MD, Vance DP, Frese SA. *et al.* GenCoF: A graphical user interface to rapidly remove human genome contaminants from metagenomic datasets. *Bioinformatics* 2019;**35**:2318–9. https://doi.org/10.1093/bioinformatics/bty963
62. Rumbavicius I, Rounge TB, Rognes T. HoCoRT: Host contamination removal tool. *BMC Bioinformatics* 2023;**24**:371. https://doi.org/10.1186/s12859-023-05492-w
63. Soverini M, Turroni S, Biagi E. *et al.* HumanMycobiomeScan: A new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. *BMC Genomics* 2019;**20**:496. https://doi.org/10.1186/s12864-019-5883-y
64. Rampelli S, Soverini M, Turroni S. *et al.* ViromeScan: A new tool for metagenomic viral community profiling. *BMC Genomics* 2016;**17**:165. https://doi.org/10.1186/s12864-016-2446-3
65. Hunt M, Swann J, Constantinides B. *et al.* ReadItAndKeep: Rapid decontamination of SARS-CoV-2 sequencing reads. *Bioinformatics* 2022;**38**:3291–3. https://doi.org/10.1093/bioinformatics/btac311
66. Hulpusch C, Rauer L, Nussbaumer T. *et al.* Benchmarking MicrobIEM - a user-friendly tool for decontamination of microbiome sequencing data. *BMC Biol* 2023;**21**:269. https://doi.org/10.1186/s12915-023-01737-5
67. Knights D, Kuczynski J, Charlson ES. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 2011;**8**:761–3. https://doi.org/10.1038/nmeth.1650
68. McGhee JJ, Rawson N, Bailey BA. *et al.* Meta-SourceTracker: Application of Bayesian source tracking to shotgun metagenomics. *PeerJ* 2020;**8**:e8783. https://doi.org/10.7717/peerj.8783
69. McKnight D, Huerlimann R, Schwarzkopf L. *et al.* microDecon: A highly accurate read-subtraction tool for the post-sequencing

removal of contamination in metabarcoding studies. *Environmental DNA* 2019;**1**:14–25. https://doi.org/10.1002/edn3.11
70. Bolyen E, Rideout JR, Dillon MR. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;**37**:852–7. https://doi.org/10.1038/s41587-019-0209-9
71. Callahan BJ, McMurdie PJ, Rosen MJ. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3. https://doi.org/10.1038/nmeth.3869
72. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol* 2019;**20**:257. https://doi.org/10.1186/s13059-019-1891-0
73. Blanco-Miguez A, Beghini F, Cumbo F. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 2023;**41**:1633–44. https://doi.org/10.1038/s41587-023-01688-w
74. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60. https://doi.org/10.1038/nmeth.3176
75. Walker MA, Pedamallu CS, Ojesina AI. *et al.* GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics* 2018;**34**:4287–9. https://doi.org/10.1093/bioinformatics/bty501
76. Borozan I, Watt S, Ferretti V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics* 2015;**31**:1396–404. https://doi.org/10.1093/bioinformatics/btv006
77. Burks DJ, Pusadkar V, Azad RK. POSMM: An efficient alignment-free metagenomic profiler that complements alignment-based profiling. *Environ Microbiome* 2023;**18**:16. https://doi.org/10.1186/s40793-023-00476-y
78. Prjibelski A, Antipov D, Meleshko D. *et al.* Using SPAdes De novo assembler. *Curr Protoc Bioinformatics* 2020;**70**:e102. https://doi.org/10.1002/cpbi.102
79. Li D, Liu CM, Luo R. *et al.* MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6. https://doi.org/10.1093/bioinformatics/btv033
80. Kang DD, Li F, Kirton E. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. https://doi.org/10.7717/peerj.7359
81. Alneberg J, Bjarnason BS, de Bruijn I. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;**11**:1144–6. https://doi.org/10.1038/nmeth.3103
82. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat Commun* 2017;**8**:2260. https://doi.org/10.1038/s41467-017-02209-5
83. van Dijk LR, Walker BJ, Straub TJ. *et al.* StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biol* 2022;**23**:74. https://doi.org/10.1186/s13059-022-02630-0
84. Vicedomini R, Quince C, Darling AE. *et al.* Strainberry: Automated strain separation in low-complexity metagenomes using long reads. *Nat Commun* 2021;**12**:4485. https://doi.org/10.1038/s41467-021-24515-9
85. Liang Q, Bible PW, Liu Y. *et al.* DeepMicrobes: Taxonomic classification for metagenomics with deep learning. *NAR Genom Bioinform* 2020;**2**:lqaa009. https://doi.org/10.1093/nargab/lqaa009
86. Mock F, Kretschmer F, Kriese A. *et al.* Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proc Natl Acad Sci U S A* 2022;**119**:e2122636119. https://doi.org/10.1073/pnas.2122636119
87. Li W, Kari L, Yu Y. *et al.* MT-MAG: Accurate and interpretable machine learning for complete or partial taxonomic assignments of metagenomeassembled genomes. *PloS One* 2023;**18**:e0283536. https://doi.org/10.1371/journal.pone.0283536
88. Shang J, Sun Y. CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. *Methods* 2021;**189**:95–103. https://doi.org/10.1016/j.ymeth.2020.05.018
89. Cres CM, Tritt A, Bouchard KE. *et al.* DL-TODA: A deep learning tool for omics data analysis. *Biomolecules* 2023;**13**:585. https://doi.org/10.3390/biom13040585
90. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* 2018;**19**:198. https://doi.org/10.1186/s13059-018-1568-0
91. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nat Commun* 2016;**7**:11257. https://doi.org/10.1038/ncomms11257
92. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. https://doi.org/10.1038/nbt.3988
93. Milanese A, Mende DR, Paoli L. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 2019;**10**:1014. https://doi.org/10.1038/s41467-019-08844-4
94. Hong C, Manimaran S, Shen Y. *et al.* PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* 2014;**2**:33. https://doi.org/10.1186/2049-2618-2-33
95. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 2015;**31**:170–7. https://doi.org/10.1093/bioinformatics/btu641
96. Roosaare M, Vaher M, Kaplinski L. *et al.* StrainSeeker: Fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ* 2017;**5**:e3353. https://doi.org/10.7717/peerj.3353
97. Zhu K, Schaffer AA, Robinson W. *et al.* Strain level microbial detection and quantification with applications to single cell metagenomics. *Nat Commun* 2022;**13**:6430. https://doi.org/10.1038/s41467-022-33869-7
98. Liao H, Ji Y, Sun Y. High-resolution strain-level microbiome composition analysis from short reads. *Microbiome* 2023;**11**:183. https://doi.org/10.1186/s40168-023-01615-w
99. Pandey S, Avuthu N, Guda C. StrainIQ: A novel n-gram-based method for taxonomic profiling of human microbiota at the strain level. *Genes (Basel)* 2023;**14**:1647. https://doi.org/10.3390/genes14081647
100. Kazantseva E, Donmez A, Frolova M. *et al.* Strainy: Phasing and assembly of strain haplotypes from long-read metagenome sequencing. *Nat Methods* 2024;**21**:2034–43. https://doi.org/10.1038/s41592-024-02424-1
101. Feng Z, Clemente JC, Wong B. *et al.* Detecting and phasing minor single-nucleotide variants from long-read sequencing data. *Nat Commun* 2021;**12**:3032. https://doi.org/10.1038/s41467-021-23289-4
102. Dilthey AT, Jain C, Koren S. *et al.* Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat Commun* 2019;**10**:3066. https://doi.org/10.1038/s41467-019-10934-2

103. Bui VK, Wei C. CDKAM: A taxonomic classification tool using discriminative k-mers and approximate matching strategies. *BMC Bioinformatics* 2020;**21**:468. https://doi.org/10.1186/s12859-020-03777-y
104. Pasolli E, Truong DT, Malik F. *et al.* Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLoS Comput Biol* 2016;**12**:e1004977. https://doi.org/10.1371/journal.pcbi.1004977
105. Oh M, Zhang L. DeepMicro: Deep representation learning for disease prediction based on microbiome data. *Sci Rep* 2020;**10**:6026. https://doi.org/10.1038/s41598-020-63159-5
106. Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–77. Long Beach, California, USA: Curran Associates Inc, 2017.
107. Ribeiro MT, Singh S, Guestrin C. "Why should I trust You?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. San Francisco, California, USA: Association for Computing Machinery, 2016.
108. Beghini F, McIver LJ, Blanco-Miguez A. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 2021;**10**:e65088. https://doi.org/10.7554/eLife.65088
109. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
110. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. *Nat Commun* 2020;**11**:3514. https://doi.org/10.1038/s41467-020-17041-7
111. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 2011;**12**:385. https://doi.org/10.1186/1471-2105-12-385
112. Letunic I, Bork P. Interactive tree of life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 2007;**23**:127–8. https://doi.org/10.1093/bioinformatics/btl529
113. Gao J, Tarcea VG, Karnovsky A. *et al.* Metscape: A Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* 2010;**26**:971–3. https://doi.org/10.1093/bioinformatics/btq048
114. Shannon P, Markiel A, Ozier O. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303
115. Dhungel E, Mreyoud Y, Gwak HJ. *et al.* MegaR: An interactive R package for rapid sample classification and phenotype prediction using metagenome profiles and machine learning. *BMC Bioinformatics* 2021;**22**:25. https://doi.org/10.1186/s12859-020-03933-4
116. Mreyoud Y, Song M, Lim J. *et al.* MegaD: Deep learning for rapid and accurate disease status prediction of metagenomic samples. *Life (Basel)* 2022;**12**:669. https://doi.org/10.3390/life12050669
117. Alshawaqfeh M, Rababah S, Hayajneh A. *et al.* MetaAnalyst: A user-friendly tool for metagenomic biomarker detection and phenotype classification. *BMC Med Res Methodol* 2022;**22**:336. https://doi.org/10.1186/s12874-022-01812-5
118. Shen WX, Liang SR, Jiang YY. *et al.* Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations. *Patterns (N Y)* 2023;**4**:100658. https://doi.org/10.1016/j.patter.2022.100658
119. Reiman D, Metwally A, Sun J. *et al.* Meta-signer: Metagenomic signature identifier based onrank aggregation of features. *F1000Res* 2021;**10**:194. https://doi.org/10.12688/f1000research.27384.1
120. Fioravanti D, Giarratano Y, Maggio V. *et al.* Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* 2018;**19**:49. https://doi.org/10.1186/s12859-018-2033-5
121. Douglas GM, Maffei VJ, Zaneveld JR. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 2020;**38**:685–8. https://doi.org/10.1038/s41587-020-0548-6
122. Langille MG, Zaneveld J, Caporaso JG. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;**31**:814–21. https://doi.org/10.1038/nbt.2676
123. Jun SR, Robeson MS, Hauser LJ. *et al.* PanFP: Pangenome-based functional profiles for microbial communities. *BMC Res Notes* 2015;**8**:479. https://doi.org/10.1186/s13104-015-1462-8
124. Erazo NG, Dutta A, Bowman JS. From microbial community structure to metabolic inference using paprica. *STAR Protoc* 2021;**2**:101005. https://doi.org/10.1016/j.xpro.2021.101005
125. Asshauer KP, Wemheuer B, Daniel R. *et al.* Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 2015;**31**:2882–4. https://doi.org/10.1093/bioinformatics/btv287
126. Wemheuer F, Taylor JA, Daniel R. *et al.* Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ Microbiome* 2020;**15**:11. https://doi.org/10.1186/s40793-020-00358-7
127. Iwai S, Weinmaier T, Schmidt BL. *et al.* Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. *PloS One* 2016;**11**:e0166104. https://doi.org/10.1371/journal.pone.0166104
128. Mongad DS, Chavan NS, Narwade NP. *et al.* MicFunPred: A conserved approach to predict functional profiles from 16S rRNA gene sequence data. *Genomics* 2021;**113**:3635–43. https://doi.org/10.1016/j.ygeno.2021.08.016
129. Bose T, Haque MM, Reddy C. *et al.* COGNIZER: A framework for functional annotation of metagenomic datasets. *PloS One* 2015;**10**:e0142102. https://doi.org/10.1371/journal.pone.0142102
130. Suzuki S, Ishida T, Ohue M. *et al.* GHOSTX: A fast sequence homology search tool for functional annotation of metagenomic data. *Methods Mol Biol* 2017;**1611**:15–25. https://doi.org/10.1007/978-1-4939-7015-5_2
131. Maranga M, Szczerbiak P, Bezshapkin V. *et al.* Comprehensive functional annotation of metagenomes and microbial genomes using a deep learning-based method. *mSystems* 2023;**8**:e0117822. https://doi.org/10.1128/msystems.01178-22
132. Cruz J, Liu Y, Liang Y. *et al.* BacMap: An up-to-date electronic atlas of annotated bacterial genomes. *Nucleic Acids Res* 2012;**40**:D599–604. https://doi.org/10.1093/nar/gkr1105
133. Chen IA, Markowitz VM, Chu K. *et al.* IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* 2017;**45**:D507–16. https://doi.org/10.1093/nar/gkw929
134. Markova N. Dysbiotic microbiota in autistic children and their mothers: Persistence of fungal and bacterial wall-deficient L-form variants in blood. *Sci Rep* 2019;**9**:13401. https://doi.org/10.1038/s41598-019-49768-9
135. Qian Y, Yang X, Xu S. *et al.* Detection of microbial 16S rRNA gene in the blood of patients with Parkinson's disease. *Front Aging Neurosci* 2018;**10**:156. https://doi.org/10.3389/fnagi.2018.00156

136. Huson DH, Auch AF, Qi J. *et al.* MEGAN analysis of metagenomic data. *Genome Res* 2007;**17**:377–86. https://doi.org/10.1101/gr.5969107
137. Lu J, Breitwieser F, Thielen P. *et al.* Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput Sci* 2017;**3**:e104. https://doi.org/10.7717/peerj-cs.104
138. Kaul A, Mandal S, Davidov O. *et al.* Analysis of microbiome data in the presence of excess zeros. *Front Microbiol* 2017;**8**:2114. https://doi.org/10.3389/fmicb.2017.02114
139. Fernandes AD, Macklaim JM, Linn TG. *et al.* ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PloS One* 2013;**8**:e67019. https://doi.org/10.1371/journal.pone.0067019
140. Jayakrishnan TT, Sangwan N, Barot SV. *et al.* Multi-omics machine learning to study host-microbiome interactions in early-onset colorectal cancer. *NPJ Precis Oncol* 2024;**8**:146. https://doi.org/10.1038/s41698-024-00647-1
141. Nyholm L, Koziol A, Marcos S. *et al.* Holo-omics: Integrated host-microbiota multi-omics for basic and applied biological research. *iScience* 2020;**23**:101414. https://doi.org/10.1016/j.isci.2020.101414
142. Zeng Y, Li J, Wei C. *et al.* mbDenoise: Microbiome data denoising using zero-inflated probabilistic principal components analysis. *Genome Biol* 2022;**23**:94. https://doi.org/10.1186/s13059-022-02657-3
143. Valous NA, Popp F, Zornig I. *et al.* Graph machine learning for integrated multi-omics analysis. *Br J Cancer* 2024;**131**:205–11. https://doi.org/10.1038/s41416-024-02706-7
144. Cheng HS, Tan SP, Wong DMK. *et al.* The blood microbiome and health: Current evidence, controversies, and challenges. *Int J Mol Sci* 2023;**24**:5633. https://doi.org/10.3390/ijms24065633
145. Wang Y, Zhao Y, Bollas A. *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;**39**:1348–65. https://doi.org/10.1038/s41587-021-01108-x
146. Wenger AM, Peluso P, Rowell WJ. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62. https://doi.org/10.1038/s41587-019-0217-9

Briefings in Bioinformatics, 2025, **26(2)**, bbaf176
https://doi.org/10.1093/bib/bbaf176
Review