Nucleic Acids Research, 2026, **54**, D763–D775
https://doi.org/10.1093/nar/gkaf1152
Advance access publication date: 8 November 2025
Database issue

HLRMDB: a comprehensive database of the human microbiome with metagenomic assembly, taxonomic classification, and functional annotation by analysis of long-read and hybrid sequencing data

Zhaoyu Zhai [1], Xiaohui Che [1], Wei Shen [2], Zishun Zhang [1], Yapeng Li [1], Jianbo Pan [1,*]

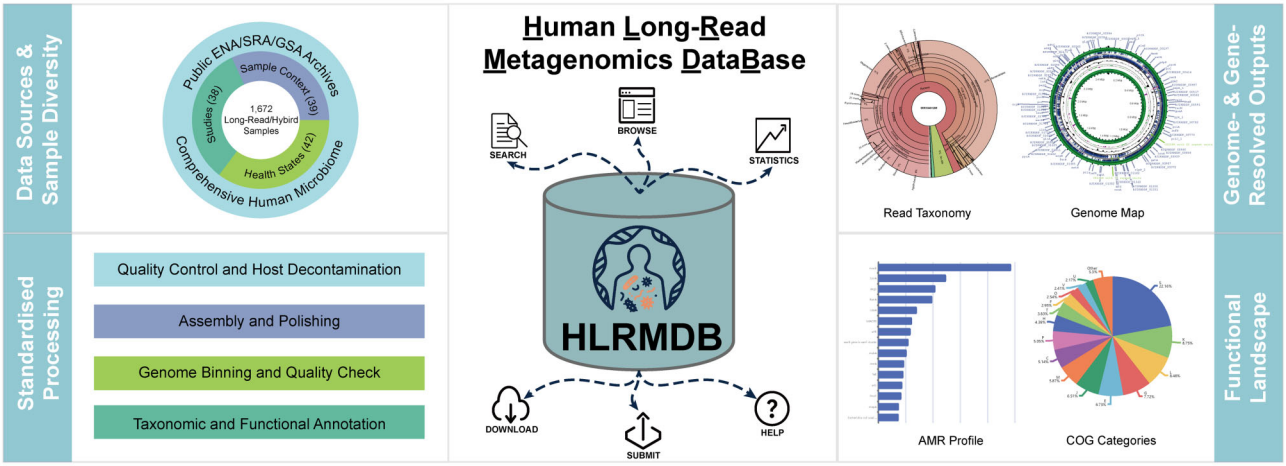
[1]Basic Medicine Research and Innovation Center for Novel Target and Therapeutic Intervention (Ministry of Education), College of Pharmacy, and Precision Medicine Center, the Second Affiliated Hospital, and Reproductive Medicine Center, the First Affiliated Hospital, Chongqing Medical University, Chongqing 400016, China
[2]Department of Infectious Diseases, Key Laboratory of Molecular Biology for Infectious Diseases (Ministry of Education), Institute for Viral Hepatitis, the Second Affiliated Hospital, Chongqing Medical University, Chongqing 400010, China

*To whom correspondence should be addressed. Email: panjianbo@cqmu.edu.cn

Abstract

The human microbiome harbours an immense diversity of uncultivated microbes; short-read metagenomic sequencing has elucidated much of this diversity, but fragment repeats and mobile elements constrain strain-level resolution. Fortunately, long-read metagenomic sequencing can generate reads spanning tens of kilobases with single-molecule accuracies exceeding 99%, enabling near-complete genome and gene cluster recovery in a cultivation-independent manner. However, systematic resources that aggregate and standardise long-read outputs remain limited. Here, we present HLRMDB (http://www.inbirg.com/hlrmdb/), a comprehensive database of human microbiome datasets derived from long-read and hybrid metagenomic sequencing. We curated 1672 publicly available metagenomes (1291 long reads; 381 hybrids) spanning 38 studies, 39 sampling contexts and 42 host health states. A uniform assembly and binning pipeline reconstructed >98 Gb of contigs and yielded 18 721 metagenome-assembled genomes (MAGs). These MAGs span 21 phyla and 1323 bacterial species, with 6339 classified as near-complete and 5609 as medium-quality. HLRMDB integrates these genome-resolved data with extensive gene-centric functional profiles and antimicrobial resistance annotations. An interactive web interface supports flexible access to both sample-level and genome-level results, with multiple visualisations linking raw reads to assembled genomes. Overall, HLRMDB offers a harmonised, long-read-oriented repository that supports reproducible, strain-resolved comparative genomics and context-sensitive ecological investigations of the human microbiome.

Graphical abstract



Received: August 11, 2025. **Revised:** October 10, 2025. **Accepted:** October 11, 2025

Introduction

The human microbiome harbours a vast diversity of microorganisms that strongly influence nutrient metabolism, metabolic homeostasis, and host immunity [1]. However, a significant fraction of these microorganisms remains recalcitrant to traditional cultivation approaches [2]. Culture-independent metagenomics now spans a spectrum of analytical scales—from direct read classification to the generation of gene-centric catalogues and, where possible, whole-genome reconstruction. Genome-resolved assemblies, often described as metagenome-assembled genomes (MAGs), have been transformative for uncovering previously unknown lineages [3]. Early successes with genome-resolved analyses demonstrated that many human-microbiome MAGs approach chromosome-level completeness, frequently retaining intact rRNA operons, plasmids, and phages [4, 5]. These resources underpin studies of strain tracking, antimicrobial resistance transmission, and community ecology [6–8]. Read- and gene-centric analyses are now indispensable for metagenomic applications that require pathogen identification, species-resolved quantification of antibiotic resistance genes, and assembly-free discovery of biosynthesis-related gene clusters, thereby extending the reach of metagenomics to samples where genome binning remains difficult [9–11].

High-throughput short-read shotgun sequencing has enabled global surveys of microbial communities [12–14]. Downstream analyses typically follow three complementary paradigms: (i) read-based classification, in which quality-filtered reads are mapped directly to reference catalogues; (ii) gene-centric assembly, in which open reading frames are predicted to construct community gene catalogues; and (iii) genome-centric assembly, in which contigs are clustered into draft genomes, plasmids or viral scaffolds for ecological inference [15, 16]. Across all of these paradigms, the inherent 150–300 bp length of short reads cannot bridge multiple kilobase repeats, leading to fragmentation of marker genes and dissociation of mobile elements or resistance loci from their microbial hosts [17–19].

Recent advances in long-read sequencing, led by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have yielded reads that can span tens to hundreds of kilobases, and PacBio HiFi has achieved ≥ 99.9% single-molecule accuracy [20–22]. While the practical read-length distributions of public datasets vary because of factors such as DNA extraction protocols and library preparation methods, these reads consistently surpass the length of short reads. When queried with k-mer classifiers or long-read aligners, these reads provide species-to-strain-level taxonomic resolution, recover low-abundance taxa and resolve structural variants invisible to short reads [23–26]. Hybrid metagenomics (combining short and long reads), which scaffolds Illumina precision onto a long-read span, increases assembly contiguity by an order of magnitude and expands community gene catalogues [27–29]. Complete biosynthesis-related gene clusters, full-length ORFs and intact resistance cassettes are now routinely assembled, while linkages between plasmids, phages and chromosomal contexts have become explicit [30, 31]. High-contiguity assemblies likewise facilitate the identification of high-quality genome bins, rRNA-complete chromosomes and genomic islands [2, 28]. Hi-C proximity ligation and native methylation patterns further refine both binning and plasmid–host associations [32, 33]. Although the volume of PacBio/ONT metagenomic data continues to increase [22], most public repositories—such as SPIRE, mOTUs-db, MGnify, and IMG/M—were conceived in the short-read era and remain dominated by Illumina-derived assemblies [34–37]. Consequently, the research community still lacks a platform that systematically archives long-read-derived polished contigs, gene catalogues and associated genome bins under a unified framework.

To address this gap, we developed HLRMDB (Human Long-Read Metagenomics Database), the first resource dedicated to long-read and hybrid human-microbiome datasets across multiple analytical levels. By prioritising long-read and hybrid assemblies, HLRMDB preserves the genomic syntax—the physical linkage of genes within operons, biosynthesis-related gene clusters (BGCs), and mobile genetic elements (MGEs)—which is often shattered in short-read repositories. This design unlocks critical research opportunities, such as tracing the horizontal transfer of complete functional modules, discovering novel natural products from intact BGCs, and, crucially, linking antibiotic resistance genes (ARGs) directly to their host genomes for accurate epidemiological surveillance. By centralising this computationally intensive work, HLRMDB provides a consistent and reproducible foundation for comparative genomics that leverages the unique advantages of long-read sequencing [38–40]. HLRMDB integrates raw and processed reads, de novo assemblies, taxonomic profiles, gene annotations and, where available, genome bins, thereby supporting functional exploration and ecological interpretation. In this study, we curated 1672 public human metagenomic samples from NCBI SRA, EMBL-EBI ENA and CNCB-NGDC GSA [41–43], covering both long-read-only and hybrid datasets. Long-read sequences were generated on the ONT or PacBio platform, while the short-read components of hybrid datasets were produced on Illumina sequencers. The collection spans diverse body sites and health conditions, capturing broad microbiome diversity. By harmonising data processing from read to gene to genome levels, HLRMDB provides an integrated foundation for studies that exploit the unique advantages of long-read sequencing—without confining users to a single output type.

Materials and methods

HLRMDB workflow

The workflow of HLRMDB is illustrated in Fig. 1. The methodology for database construction and the guidelines for using the web server are described below.

Data collection and curation

We conducted a comprehensive PubMed search using the keywords '"Nanopore" AND "metagenomics" and "PacBio" AND "metagenomics"' to capture all long-read or hybrid metagenomic studies published before May 2025. Titles and abstracts were manually screened, and full texts were consulted where necessary to retain only investigations that analysed human-derived microbiome samples. Studies based on mock communities, *in silico* simulations, or purely methodological benchmarks were excluded. For each eligible article, we first extracted the BioProject accession(s) and sequenced run identifiers. We then cross-referenced these identifiers with metadata programmatically retrieved from NCBI SRA [41], EMBL-EBI ENA [42], and CNCB-NGDC GSA [43] to classify

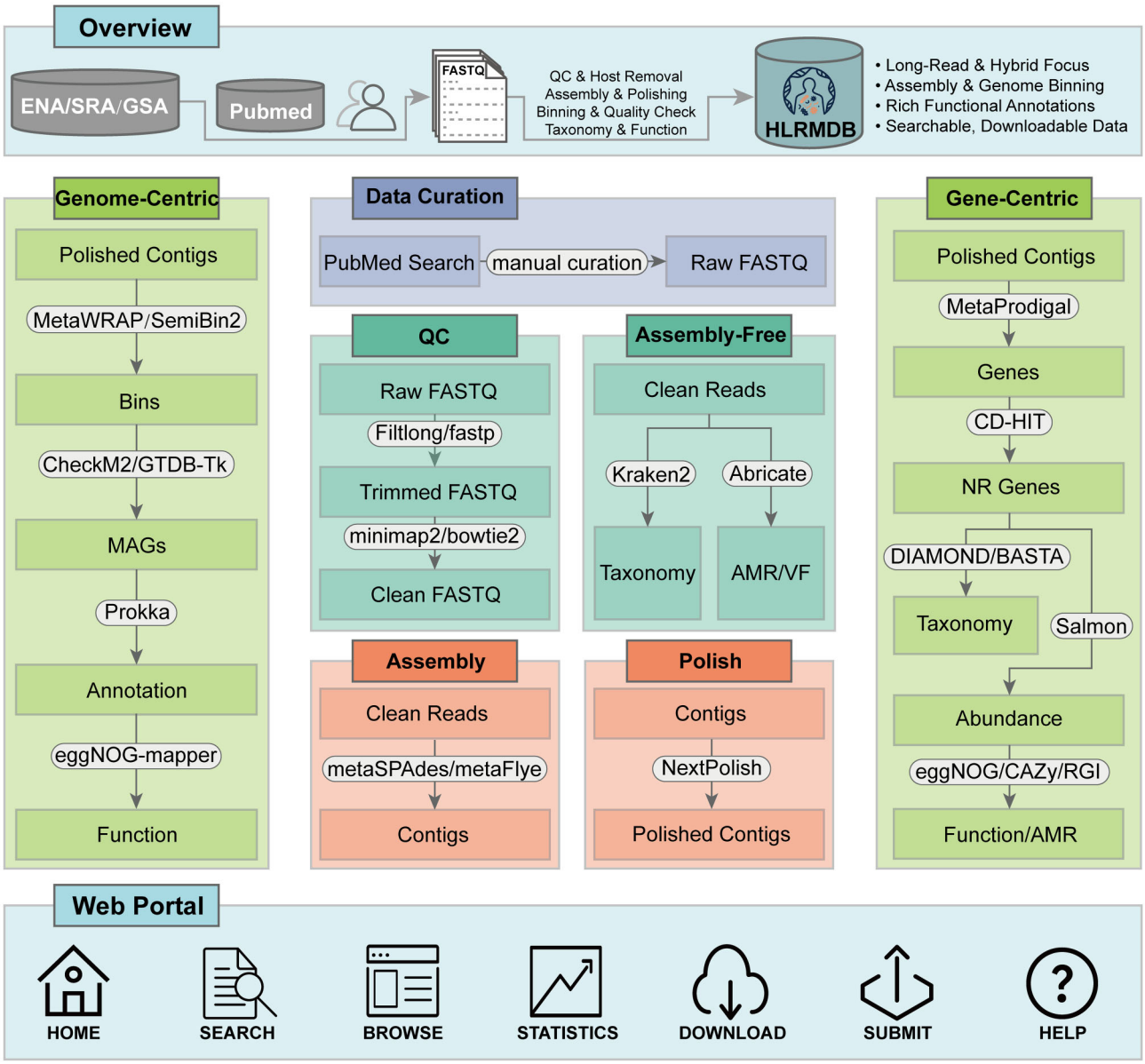


Figure 1. HLRMDB workflow and data types. Human long-read and hybrid metagenomes were identified through a PubMed-guided literature curation and retrieved from ENA, SRA, and GSA. Host-depleted reads undergo quality filtering and human-sequence removal. Two analysis modules are then applied: (i) an assembly-free read-level module for taxonomic profiling and detection/quantification of antimicrobial resistance and virulence markers, and (ii) an assembly-based module comprising *de novo* assembly, polishing, binning, quality and taxonomic assessment, and functional annotation. All the results are deposited into an interactive database supporting exploration via linked tables and visualisations.

every sample as either long-read only (ONT or PacBio) or hybrid (Illumina paired with ONT/PacBio). Additional sample-level descriptors—the sequencing platform, sampling context, and host health status—were curated into a unified metadata table. Raw FASTQ files were downloaded with IBM Aspera CLI and checked for integrity via MD5 digestion before downstream analysis.

Sequencing data preprocessing

Raw long-read FASTQ archives were initially inspected for file-format compliance and unexpected truncations; any reads failing these integrity checks were discarded before further processing. The remaining reads were subjected to quality control with Filtlong (v0.2.1), which preferentially retains higher-quality, longer fragments while trimming low-confidence regions [44]. Postfilter metrics were visualised and exported as interactive HTML files using NanoPlot (v1.42.0) [45], providing an auditable record of dataset quality. To remove human-derived sequences, the filtered long reads were aligned against the current human reference genome with minimap2 (v2.24) [46]. We parsed the alignment results and retained only unmapped reads for downstream analysis. Adapters and low-quality trailing bases were excised with fastp (v0.23.1), which also produced comprehensive pre/posttrim quality summaries [47]. The cleaned pairs were then mapped to the same human reference using Bowtie2 (v2.3.5.1) [48], and all non-aligned read pairs were extracted with SeqKit (v2.9.0) [49]. The resulting host-depleted short- and long-read sets were subjected to subsequent assembly, binning, and annotation stages.

Assembly-free read-level analysis

Preprocessed reads from each sample were directly classified taxonomically using Kraken2 (v2.1.3) [50] against its standard database (September 2024), generating compositional profiles at the species level. The resulting Kraken2 reports were visualised as interactive hierarchical charts with KronaTools (v2.8.1) [51] to facilitate the exploration of community composition. For resistome (antibiotic resistance) and virulome profiling, we employed Abricate (v1.0.1) to screen the NCBI AMRFinderPlus [52] and VFDB [53] databases, retaining all hits that met default criteria. The functional gene abundances were quantified as gene copies per gigabase (gc/Gb) of sequencing data using the 'abundance_calculate.py' script from EasyNanoMeta [54].

Hybrid and long-read genome assembly and polishing

Hybrid datasets—combining Illumina short reads with their corresponding ONT or PacBio long reads—were de novo assembled using metaSPAdes (v3.13.0) [55], leveraging the complementary strengths of short and long reads. Long-read-only datasets were assembled using metaFlye (v2.9) [56] to exploit the ability of long reads to resolve repeats and recover highly contiguous contigs. All the assemblies were subjected to an initial round of long-read self-polishing with NextPolish (v1.4.1) [57]. The hybrid assemblies then received a second polishing step using their short-read data to further improve the base-level accuracy. The polished assemblies were evaluated with metaQUAST (v5.3.0) [58], which provided comprehensive metrics of assembly performance for each sample.

Gene prediction, annotation, and abundance profiling in hybrid assemblies

Genes were predicted from assembled contigs using MetaProdigal (v2.6.3) [59]. The resulting nucleotide sequences were dereplicated with CD-HIT (v4.8.1) [60], and representative sequences from each cluster were translated into protein sequences using SeqKit (v2.9.0) [49]. The nonredundant protein set was then taxonomically annotated using DIAMOND (v0.9.21) [61] against the NCBI NR database [41]. Species-level assignments were refined with BASTA (v1.4) [62], and the results were visualised using KronaTools. Salmon (v0.13.1) [63] was used to quantify the gene cluster abundances of the hybrid assemblies. Proteins were further functionally annotated with eggNOG-mapper (v2.1.12) [64] against the eggNOG 5.0 database [65]. CAZy (carbohydrate-active enzyme) families were identified by DIAMOND searches [61] against the CAZy database [66], and potential antibiotic resistance genes were detected using the Resistance Gene Identifier (RGI) in conjunction with the CARD repository [67]. Finally, summary scripts from EasyMetagenome [68] (e.g. 'summarizeAbundance.py') were used to consolidate KEGG orthologue profiles, COG functional categories, CAZy family counts, and resistance gene abundances for each sample.

Genome binning, quality assessment, and taxonomic classification

The long-read assemblies were binned into draft genomes using SemiBin2 (v2.1.0) [69] in long-read mode. Hybrid assemblies were binned with MetaWRAP (v1.3.2) [70], running multiple binning algorithms (MetaBAT2, MaxBin2, and CONCOCT) and incorporating SemiBin2 results. The candidate bins from these methods were dereplicated and refined using the DAS Tool (v1.1.7) [71] to produce a nonredundant set of high-quality genome bins. Bin quality (completeness and contamination) was evaluated with CheckM2 (v1.0.2) [72], and species-level taxonomy was assigned using GTDB-Tk (v2.4.0) [73]. We assigned each bin a quality tier on the basis of the MIMAG standard [4]. Bins with ≥90% completeness and ≤5% contamination are classified as 'near-complete' [5]. We use this term in lieu of 'high-quality' because the formal MIMAG definition additionally requires rRNA operons (5S, 16S, and 23S) and ≥18 tRNAs, which were not screened for in this release. Bins with ≥50% completeness and <10% contamination are labelled 'medium-quality', and the remainder are 'low-quality' drafts. Each recovered bin was functionally annotated using Prokka (v1.14.6) [74] for gene prediction, and its proteins were further analysed with eggNOG-mapper (v2.1.12) [64] to assign COG and pathway annotations. Circular genome maps of bins were generated using CGView (v2.0.3) [75] for visualisation.

Web Interface Implementation

The HLRMDB platform runs on a Linux server configured with Nginx (v1.14.1) as the web server and uWSGI (v2.0.20) as the application server; all metadata and analysis results are stored in a MySQL relational database (v8.0.26). The backend is implemented with Django (v2.1.8), while the frontend leverages Bootstrap (v4.3.1), jQuery (v3.2.1), DataTables (v1.10.21), and ECharts (v5.0.2) to deliver dynamic, interactive visualisations. Analytical pipelines and statistical computations are executed in Python, employing libraries such as NumPy (v1.21.2) and pandas (v1.3.5).

Results

Database content and statistics

HLRMDB currently aggregates 1672 publicly available human metagenomic samples. This collection includes 1291 datasets sequenced exclusively with long-read technologies (1290 ONT; 1 PacBio) and 381 generated by hybrid strategies pairing Illumina short reads with long reads (357 ONT-hybrid; 24 PacBio-hybrid). Where chemistry/version metadata were available, ONT datasets were dominated by R9.4.1 (n = 532) and R9.4 (n = 298), with a smaller R10.4.1 / Kit14 subset (n = 33) and legacy R7.3 (n = 6). PacBio datasets were mainly Sequel Chemistry v3.0 (n = 23) with rare RS II P6-C4 (n = 2). These samples were drawn from 38 independent studies and span 39 distinct sampling contexts (e.g. faeces, skin, respiratory tract) as well as 42 host health states. Assembly of all datasets using a uniform pipeline produced >98 Gb of assembled contigs in total, with the longest single contig extending 9.5 Mb. Genome binning of the assemblies yielded 18 721 draft genomes (MAGs/bins). Following the MIMAG standard, we classified these bins into quality tiers. Among the total bins, 6339 meet near-complete quality standards (completeness ≥ 90%, contamination ≤ 5%), and an additional 5609 are of medium quality (completeness ≥ 50%, contamination < 10%). The remaining 6 773 bins are classified as low-quality drafts. Taxonomic classification places these genomes into 21 phyla comprising 1323 bacterial species, substantially expanding the long-read-derived ge-

nomic repertoire of the human microbiome. At the sample level, read-length distributions vary widely, although a substantial fraction of raw long reads extend for several kilobases and can traverse repeat-rich regions. To facilitate quality assessment, we report the read-length N50 and the fraction of reads <1 kb for each sample (Supplementary Table S1). Additionally, platform- and chemistry-aware boxplots (median with interquartile range, with markers for the 95th percentile and maximum) show that hybrid assemblies often outperform long-read-only assemblies on contiguity and accuracy metrics, consistent with the complementary strengths of combining short and long reads. The current dataset is predominantly derived from Oxford Nanopore Technologies (ONT), with a smaller subset from Pacific Biosciences (PacBio). Where provided in the source metadata, we record and expose the specific sequencing chemistry or version.

Web design and interface

HLRMDB delivers flexible data exploration through an intuitive web interface (Fig. 2). Four complementary entry points—Quick Search, Advanced Search, Browse, and Download—allow users to retrieve data either at the sample level or as individual genome bins, according to various needs (Fig. 2A–C). Selecting a sample from any results table opens a detail page that provides a comprehensive overview of the long-read or hybrid dataset and the associated downstream analyses (Fig. 3). Clicking a specific bin on a sample's detail page directs the user to a dedicated bin-detail page that displays rich annotation metadata and interactive visualisations for that genome bin (Fig. 4).

All public dashboard summaries are tied to the same frozen release used for the manuscript analyses, with subsequent releases versioned and selectable via a release switch. The statistics page presents distribution-aware views (median with interquartile range, with markers for the 95th percentile and maximum) across platform-stratified panels, ensuring that published and displayed values remain reproducible while allowing users to explore later releases transparently (Fig. 2D). The download page supports bulk retrieval of all processed sample datasets and genome bins and provides source code for metagenomics data analysis (Fig. 2C). To enable user-side execution with data privacy and reproducibility, we will release the full HLRMDB workflow as a containerised pipeline (Docker/Singularity) suitable for local, HPC, or cloud environments; this toolkit mirrors the server-side processing used for public releases and provides a pragmatic bridge towards a future web-based analysis service. To facilitate robust comparative analyses, users can download either the complete set of all MAGs or a prefiltered subset containing only near-complete and medium-quality genomes. Community submissions are enabled via a Submission page for contributing new datasets spanning diverse health states, and a Help page provides step-by-step tutorials and contact information for users. This curated submission model ensures that all the data integrated into HLRMDB adhere to a consistent quality and processing standard, maintaining the integrity of the harmonised resource. To improve data exploration, the Advanced Search and Browse tables include a 'Sequencing platform' filter (with specific chemistry/version facets when available), and the statistics page offers platform-stratified panels so that users can filter and compare aggregate metrics (yield, read-length N50, assembly contiguity, and MAG quality) for ONT vs. PacBio data. Furthermore, the MAG tables (Browse and Sample views) now include a 'quality_tier' badge for each genome bin, with low-quality bins clearly flagged visually. The download page offers two dataset options: the full set of MAGs and a default high-quality subset (near-complete + medium-quality bins) intended for convenient comparative analyses.

Data searching

HLRMDB offers two complementary search modes to suit different user needs (Fig. 2A). Quick search, available from a homepage widget, has separate 'Sample' and 'Genome Bin' tabs: in Sample mode, free-text keywords for host health/disease (e.g. healthy), sampling context (e.g. stool), or sequencing platform (e.g. Nanopore) immediately retrieve matching datasets, whereas Genome Bin mode accepts bin identifiers or taxonomic terms (e.g. Bacilli) to locate specific genome bins. For finer control, the Advanced Search interface exposes comprehensive metadata filters—Sample datasets can be narrowed by BioProject ID, sampling context, host health status, sequencing platform, or read count, and now include read-quality filters: read-length N50 with presets ($\geq$5, $\geq$10, and $\geq$20 kb) and Reads <1 kb (%). Genome bins can be filtered by sample ID, quality metrics (e.g. completeness, contamination, contig N50), or taxonomic rank (e.g. phylum). Sample-level filters include a sequencing platform and, when present in the metadata, the specific sequencing chemistry or version, enabling users to perform platform-aware dataset selection. Both search modes return results in interactive, sortable tables that users can further investigate.

Data browsing

Within the Browse section of HLRMDB (Fig. 2B), users interact with the resource through two coordinated views. The Sample Datasets view presents an interactive table of every long-read or hybrid sample, offering real-time filters for the BioProject, sampling context, host health/disease status and other metadata; thus, datasets can be swiftly subset to match specific study criteria. The Genome bins view displays all recovered MAGs/bins in a hierarchical taxonomic tree generated from GTDB-Tk annotations; expanding any node—from phylum to species—executes a query that returns a table of all bins assigned to that lineage, whereas entries lacking confident placement at that rank appear under an 'Unclassified' category. Together, these linked views allow users to locate taxa at multiple phylogenetic depths and seamlessly retrieve the corresponding MAG/bin information for downstream analyses.

Detailed Information

Clicking a Sample ID in any Search or Browse result opens a multipanel detail page (Fig. 3) that walks users from raw reads to genome-resolved insights for that specimen. The page begins with Sample metadata and QC, summarising accession identifiers, sequencing strategy, sampling context and host health/disease alongside interactive quality−control plots that juxtapose long- and short-read metrics (read−length distributions, base−quality scores) for immediate visual assessment (Fig. 3A and B). A Read-level composition module then provides an assembly-free overview that facilitates rapid sample-level assessment: a sunburst chart depicts the taxonomic profiles, while bar and pie charts show the antibiotic resistance

Figure 2. Contents of HLRMDB. (**A**) Search section: a quick search widget and an advanced search form enable free-text queries and faceted filtering; matches are returned in sortable result tables for immediate inspection. (**B**) Browse section: an interactive sample table supports filters such as BioProject, sampling context, host health status and sequencing platform; a GTDB-based taxonomy tree retrieves lineage-specific genome-bin tables. (**C**) Download & Submit section. (**D**) Statistics section: an interactive dashboard summarises database-wide content, including read-length distributions, metrics distributions, and a heatmap of completeness and contamination across genome bins.

categories, individual resistance genes and virulence factors detected directly in the reads (Fig. 3C). The Assembly and Genome Bins section follows, integrating contig- and genome-level outputs by presenting an assembly report (N50, contig count, and GC content) together with a sortable table of recovered bins annotated for completeness, contamination and taxonomy; interactive scatter plots, bar charts and a phylogenomic tree place each bin within the reference context and highlight taxonomic placement (Fig. 3D). Next, the gene-centric analysis module visualises ORF length, GC content, start codon usage and coding sequence completeness and summarises CAZy families, COG categories and KEGG pathways across hierarchical levels (Fig. 3D). A dedicated Resistome dashboard further dissects antimicrobial resistance genes, displaying overall abundance and breakdowns by gene family, drug class and mechanism (Fig. 3D). Finally, each bin links to a Bin detail view that recaps quality metrics and taxonomy, shows an interactive circular genome map of contigs and rRNA operons, and replicates the functional and resistome visualisations for that individual genome (Fig. 4A–C). Together, these multiscale modules allow users to move seamlessly from community-level summaries to assembly-free and

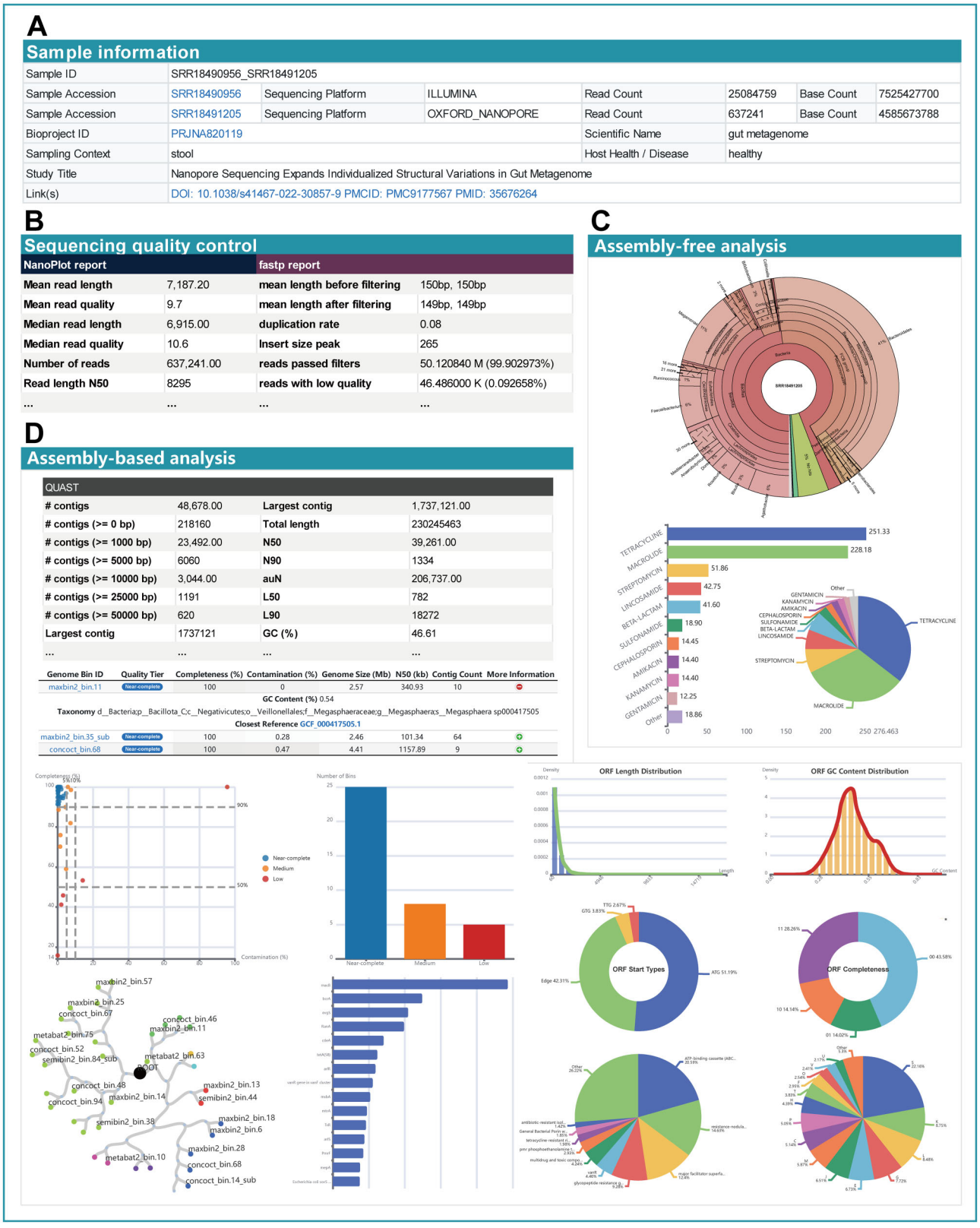


Figure 3. Detailed information for an HLRMDB sample. (**A**) Sample information: consolidated specimen metadata and study links for rapid context. (**B**) Sequencing quality control: paired long- and short-read quality summaries for preanalysis screening. (**C**) Assembly-free analysis: read-level taxonomic composition with overviews of antimicrobial resistance and virulence markers, providing an immediate, assembly-independent snapshot. (**D**) Assembly-based analysis: compact assembly summaries linked to genome-bin exploration and reference-context visualisations, complemented by gene-centric functional and resistome profiles.

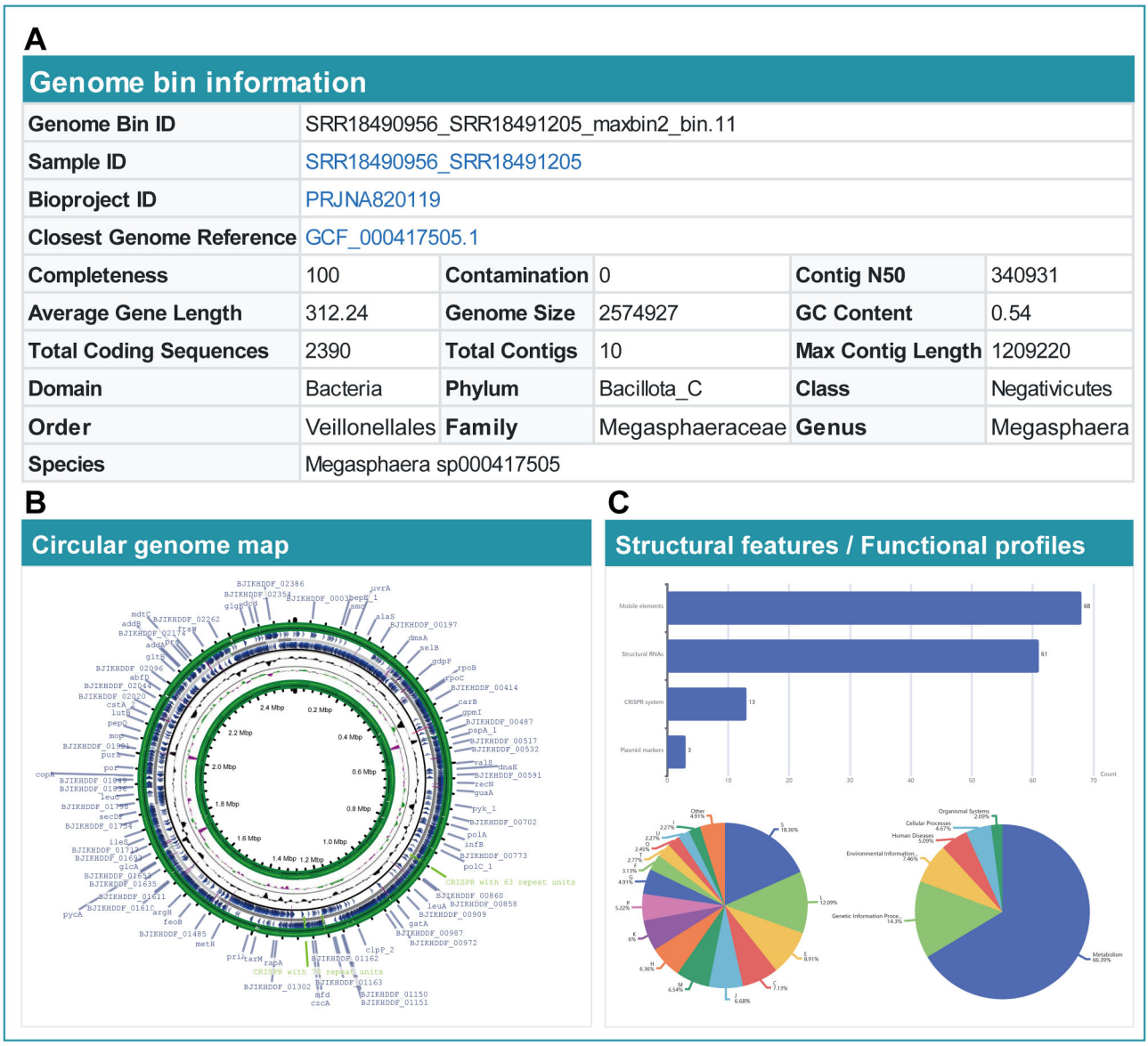


Figure 4. Bin-level contents within HLRMDB. (**A**) Genome bin information: consolidated identifiers and links with key quality metrics and taxonomy for rapid assessment. (**B**) Circular genome map: interactive overview of contigs, coding features and rRNA operons for structural context. (**C**) Structural features and functional profiles: summaries of structural modules and aggregated functional classifications for the individual genome.

assembly-based analyses, maximising the interpretability of long-read and hybrid metagenomic datasets within HLRMDB.

Case study

To illustrate the analytical value of HLRMDB in a disease-focused context, we reexamined a human faecal metagenome (ERR6752166_ERR6713810) from an obese donor that had been sequenced using a hybrid (short- and long-read) strategy. Using the Browse page filters, we identified a sample tagged with the host phenotype 'obese'. The read-level taxonomic profile of this sample indicated that *Prevotella copri* contributed ~40% of all classifiable long reads. In animal models, high *P. copri* abundance is positively associated with obesity [76] and adipose tissue deposition [77], yet different *P. copri* strains can have markedly heterogeneous metabolic effects [78]. Rapid resistome screening of the sample revealed high-coverage reads for the genes *cfxA6*, *ermF*, and *tetQ*/*tetX*. These genes frequently occur together as a collinear cassette of mobile elements in *Prevotella* spp. [79], suggesting that historical antibiotic pressure may have synergised with the host's metabolic dysregulation to reshape the gut community in this individual.

The hybrid assembly of the sample's reads recovered a near-complete *P. copri* genome (ERR6752166_ERR6713810_concoct_bin.23; MAG completeness ≈ 100%, contamination ≈ 4%). Within a single 261 kb contig of this genome, we identified multiple polysaccharide utilisation loci (PULs), which shed light on the metabolic potential of the strain (Supplementary Table S2). One PUL gene encodes the canonical SusC/SusD transporter coupled with GH13/GH97 glycosidases for starch and α-glucan catabolism [80]; another PUL gene is highly homologous to a pectin side chain degradation system [81]. Functionally equivalent PUL systems have been experimentally characterised

across multiple *P. copri* lineages [82], suggesting that this strain is capable of liberating fermentable oligosaccharides from dietary polysaccharides. These oligosaccharides are subsequently converted by the microbiota into short-chain fatty acids (SCFAs), which fuel the host's energy metabolism [83] and are correlated with obesity phenotypes in human cohorts [84]. Notably, the same contig in this *P. copri* genome harboured an IS21-like insertion sequence and an XerC site-specific recombinase, providing a genetic mechanism for PUL acquisition and horizontal transfer [85]. In longitudinal dietary intervention studies, such XerC-mediated insertions have been observed to fluctuate rapidly in the gut microbiomes of obese individuals [86], suggesting that mobile genetic elements further modulate carbohydrate metabolic potential during obesity. In this example, leveraging long reads—alone or in cost-effective hybrid assemblies—elevated contig continuity by an order of magnitude, recruited virtually all reads, and preserved the physical linkage between PULs and resistance cassettes. These advantages revealed functional and epidemiological insights that would remain inaccessible from short-read metagenomes alone.

Conclusions and future directions

Long-read sequencing and cost-efficient long/short-read hybrid protocols are steadily reshaping metagenomic research. These technologies yield assemblies that are typically several-fold more contiguous and complete than those from short-read data alone, enabling finer strain discrimination and more reliable linkage of antimicrobial resistance or mobile element loci to their host genomes. HLRMDB capitalises on this long-read contiguity at scale, providing the research community with a unique resource to investigate the interplay between genomic organisation and microbial ecology. Three key attributes distinguish HLRMDB from existing microbiome repositories. First, its prioritisation of long-read and hybrid data preserves the genomic context—the 'genomic syntax'—essential for resolving the colocalisation of functional elements. Second, an interactive multiscale interface links data across the read, assembly, and genome levels. The sample-level interface unifies assembly metrics, binning outcomes and functional summaries in a single view, enabling one-click pivots across metrics while preserving filters and provenance. Third, by executing a standardised, computationally intensive processing pipeline upfront, HLRMDB democratises access to high-resolution genomic data, offering a harmonised and reliable starting point for downstream investigations.

To date, the human microbiome field has lacked a dedicated platform to organise the rapidly growing body of long-read datasets and their genome-resolved outputs. To situate HLRMDB within the current landscape, we provide a structured comparison with representative resources (Supplementary Table S3), distinguishing human gut short-read MAG catalogues (e.g. UHGG, HRGM)[87, 88], cross-biome consistently processed resources (e.g. SPIRE) [34], general metagenomic platforms (e.g. MGnify, IMG/M) [36, 37], and marker-gene resources (e.g. mOTUs-db) [35]. The fundamental limitation of short-read-centric architectures is the systematic fragmentation of microbial genomes, which results in the dissociation of mobile genetic elements from their hosts and the disruption of large functional gene clusters. HLRMDB is designed to address this gap by integrating long-read and hybrid metagenomes under a uniform pipeline that preserves locus-level genomic context for genome-resolved analyses. For example, the cross-biome, consistently processed resource SPIRE encompasses 99 146 metagenomic samples and 1.16 million newly constructed MAGs and was explicitly built from ENA whole-genome shotgun datasets sequenced on the Illumina platform, offering planetary-scale taxonomic breadth rather than a human-specific, long-read-native catalogue; assemblies are therefore typically short-read-derived and more fragmented than long-read or hybrid counterparts. Similarly, the mOTUs online database (mOTUs-db) provides web-accessible genomic context and comprises 2.83 million MAGs reconstructed de novo from 117 902 metagenomic samples, but it is designed to contextualise marker-gene profiles rather than to deliver a uniform long-read-aware assembly pipeline. General-purpose platforms such as MGnify (v5.0) support amplicon, reads and assembly workflows, with long-read or hybrid assembly available via a request-based mechanism, yet processing is not uniformly applied across legacy submissions. In contrast, HLRMDB is currently the first repository devoted to systematically processing and harmonising long-read and hybrid metagenomic data from the human microbiome, providing a high-contiguity resource that preserves the genomic context intractable with short-read data alone.

The current release is built predominantly on public data from ONT platforms, and a clear understanding of the intrinsic differences between technologies is vital for users to select appropriate datasets and correctly interpret results. Long–read sequencing platforms differ markedly in read length, accuracy and error patterns, and these differences influence downstream analyses. PacBio's single–molecule real–time (SMRT) platform generates circular consensus (HiFi) reads with error rates below 1%. Its polymerase–limited SMRT chemistry typically produces reads of 10–30 kb (insert sizes 1–100 kb), and errors are largely random indels [89]. In contrast, ONT offers the longest reads—often 10–30 kb and occasionally exceeding 2.3 Mb—but raw error rates are historically higher (≤5% with current Q20/Q30 chemistries [20, 90]). ONT errors cluster in homopolymeric regions, necessitating consensus or duplex base–calling to improve accuracy. Recent pharmacogenomics reviews highlight that advances in base–calling algorithms and chemistry have lowered raw ONT and PacBio error rates to <5% and <1%, respectively [89]. Owing to these differences, PacBio is often preferred for single–nucleotide variant detection and haplotype phasing, whereas ONT's ultra–long reads enable assembly across repetitive or structural variant–rich regions. Hybrid strategies exploit the complementary strengths of short–read and long–read data: inexpensive PacBio CLR or standard ONT reads (with error rates 5–15%) can be corrected using accurate Illumina reads, and the more costly PacBio HiFi or ONT Q20 + data (∼1% error) can be combined with short reads to yield high–accuracy assemblies. Methodological work shows that hybrid corrected reads are even more accurate than the latest HiFi or ONT reads alone, because the complementary length and accuracy of third–generation and next–generation reads produce assemblies that are both long and reliable [91]. Consequently, researchers must tailor variant–calling and assembly pipelines to account for platform–specific error modes and may adopt hybrid correction or joint assemblies to achieve the desired balance between read length, cost and accuracy. Recognising this challenge, the HLRMDB workflow incorporates multiple rounds of assembly polishing specifically de-

signed to correct indel errors and improve base-level accuracy, including short-read polishing where available [92], thereby enhancing the reliability of gene prediction and functional annotation for all datasets, including those generated with earlier ONT chemistries.

Additionally, the wide variation in read lengths observed in public datasets is the result of a combination of upstream sample handling and downstream processing choices. The primary determinant of read length is the integrity of the input high-molecular-weight (HMW) DNA. Achieving reads in the 10–100 kb range is contingent upon specialised extraction methods, while attaining ultra-long reads (>100 kb) demands even more meticulous protocols [93]. However, the prevalence of shorter fragments in many datasets is largely attributable to upstream sample preparation. Commercial DNA extraction kits, often optimised for short-read sequencing, frequently employ vigorous methods like bead-beating that inevitably shear HMW DNA [94]. Consequently, preserving DNA integrity necessitates a renewed focus on gentler, albeit more laborious, techniques. This introduces a critical trade-off, as procedural modifications aimed at preserving long fragments—such as using wide-bore tips and minimising agitation—often reduce total DNA yield. Similarly, size-selection steps designed to eliminate short fragments may be omitted to avoid significant sample loss, which can reach as high as 40% [95]. These upstream challenges are further compounded by downstream choices in library preparation and sequencing. Ligation-based library protocols are generally preferred for maintaining fragment length, whereas strategies involving PCR amplification or tagmentation inherently curtail read length due to polymerase processivity limits or enzymatic fragmentation, respectively [96]. Furthermore, the evolution of sequencing chemistries significantly influences the final read-length profile. For instance, a comparison of Oxford Nanopore R9.4.1 and R10.4 flow cells revealed divergent read-length distributions from identical DNA, yielding median lengths of 6.3 and 2.9 kb, respectively [97]. This highlights that even recent chemistries can routinely produce reads well below 10 kb. In summary, the observed read-length distribution is the cumulative outcome of this multi-stage process. The abundance of short reads is a predictable consequence of using standard extraction kits without stringent size selection, further shaped by library strategy and sequencing chemistry. Therefore, achieving consistently long reads requires a holistic optimisation of the entire workflow, from initial sample handling through to final data analysis.

The present HLRMDB release aggregates 1672 public samples (1291 long-read and 381 hybrid) spanning 38 studies, 39 sampling contexts, and 42 host health states. A uniform assembly-to-binning workflow recovered 18 721 genome bins from these datasets, of which 6339 met near-complete quality standards; collectively, all these MAGs/bins represented 1323 species across 21 bacterial phyla. Three key attributes distinguish HLRMDB from existing microbiome repositories. First, prioritisation of long-read and hybrid data preserves the genomic context—complete rRNA operons, mobile elements, and multigene cassettes—which are often fragmented into short-read assemblies. This enables more accurate strain-resolved ecological analyses and comprehensive resistome profiling. Second, an interactive multiscale interface links data at the read, assembly, and genome levels, allowing users to smoothly transition from community-wide summaries to detailed single-genome pages. This integrated platform facilitates exploration across multiple taxonomic and functional scales. Third, a standardised, computationally intensive processing pipeline is executed upfront, providing users with uniformly curated datasets. By centralising the labour- and resource-intensive steps of long-read metagenomic analysis, HLRMDB saves researchers from the need to perform de novo assembly and binning. This approach democratises access to high-resolution genomic data, offering an immediate and reliable starting point for downstream biological investigations for a broad range of researchers, regardless of their access to high-performance computing resources.

Moving forward, HLRMDB will be updated on a quarterly basis, continuously incorporating new long-read metagenomic datasets and incrementally expanding its analytical scope. To further empower the community and promote reproducible science, our immediate priority is to release the complete HLRMDB analysis workflow as a containerised pipeline (Docker/Singularity). This will provide a sustainable, scalable, and secure solution for researchers to apply our validated methods to their own datasets. Importantly, this portable workflow is a foundational step, creating the necessary engine that could power a future web-based analysis service, thereby aligning our near-term deliverables with the long-term vision of a fully interactive analytical platform.

In summary, HLRMDB consolidates diverse long-read metagenomic datasets into a coherent, searchable repository. Coupling a genome-resolved context with interactive visualisation can support ongoing investigations into microbiome ecology, evolution, and host interactions, providing a practical platform for both experimental and computational researchers.

Acknowledgements

The computing work in this paper was partly supported by the Supercomputing Center of Chongqing Medical University.

Author contributions: Z.Z. (Data curation [lead], Formal Analysis [lead], Investigation [equal], Methodology [lead], Resources [lead], Writing – original draft [lead], Writing – review & editing [equal], X.C. (Data curation [equal], Formal Analysis [supporting], W.S. (Methodology [supporting], Writing – review & editing [supporting], Z.Z. (Data curation [supporting], Y.L. (Data curation [supporting], J.P. (Conceptualisation [lead], Funding acquisition [lead], Investigation [equal], Methodology [equal], Project administration [lead], Supervision [lead], Writing – review & editing [equal].

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Natural Science Foundation of China (32470699). Funding to pay the Open Access publication charges for this article was provided by Chongqing Medical University.

Data availability

HLRMDB is a freely accessible online database available at http://www.inbirg.com/hlrmdb/. The source code of HLRMDB for metagenomics data analysis is available at http://www.inbirg.com/hlrmdb/downloads/.

References

1. Hou K, Wu Z-X, Chen X-Y *et al.* Microbiota in health and diseases. *Sig Transduct Target Ther* 2022;7:135. https://doi.org/10.1038/s41392-022-00974-4
2. Kim CY, Ma J, Lee I HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat Commun* 2022;13:6367. https://doi.org/10.1038/s41467-022-34149-0
3. Kim N, Ma J, Kim W *et al.* Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp Mol Med* 2024;56:1501–12. https://doi.org/10.1038/s12276-024-01262-7
4. The Genome Standards Consortium, Bowers RM, Kyrpides NC *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;35:725–31. https://doi.org/10.1038/nbt.3893
5. Benoit G, Raguideau S, James R *et al.* High-quality metagenome assembly from long accurate reads with metaMDBG. *Nat Biotechnol* 2024;42:1378–83. https://doi.org/10.1038/s41587-023-01983-6
6. Zachariasen T, Russel J, Petersen C *et al.* MAGinator enables accurate profiling of de novo MAGs with strain-level phylogenies. *Nat Commun* 2024;15:5734. https://doi.org/10.1038/s41467-024-49958-8
7. Baek JW, Lim S, Park N *et al.* Extensively acquired antimicrobial-resistant bacteria restructure the individual microbial community in post-antibiotic conditions. *npj Biofilms Microbiomes* 2025;11:78. https://doi.org/10.1038/s41522-025-00705-x
8. Eisenhofer R, Odriozola I, Alberdi A Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun* 2023; 3:12. https://doi.org/10.1038/s43705-023-00221-z
9. Alcolea-Medina A, Alder C, Snell LB *et al.* Unified metagenomic method for rapid detection of microorganisms in clinical samples. *Commun Med* 2024;4:135. https://doi.org/10.1038/s43856-024-00554-3
10. Chen X, Yin X, Xu X *et al.* Species-resolved profiling of antibiotic resistance genes in complex metagenomes through long-read overlapping with Argo. *Nat Commun* 2025;16:1744. https://doi.org/10.1038/s41467-025-57088-y
11. Gupta VK, Bakshi U, Chang D *et al.* TaxiBGC: a taxonomy-guided approach for profiling experimentally characterized microbial biosynthetic gene clusters and secondary metabolite production potential in metagenomes. *Msystems* 2022;7:e0092522. https://doi.org/10.1128/msystems.00925-22
12. Ordonez A, Hussain U, Cambon MC *et al.* Evaluating agar-plating and dilution-to-extinction isolation methods for generating oak-associated microbial culture collections. *ISME Commun* 2025; 5:ycaf019. https://doi.org/10.1093/ismeco/ycaf019
13. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–14. https://doi.org/10.1038/nature11234
14. Thompson LR, Sanders JG, McDonald D *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;551:457–63. https://doi.org/10.1038/nature24621
15. Soufi HH, Porch R, Korchagina MV *et al.* Taxonomic variability and functional stability across Oregon coastal subsurface microbiomes. *Commun Biol* 2024;7:1663. https://doi.org/10.1038/s42003-024-07384-y
16. Legrand TPRA, Alexandre PA, Wilson A *et al.* Genome-centric metagenomics reveals uncharacterised microbiomes in Angus cattle. *Sci Data* 2025;12:547. https://doi.org/10.1038/s41597-025-04919-8
17. De Maio N, Shaw LP, Hubbard A *et al.* Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb Genomics* 2019;5:e000294. https://doi.org/10.1099/mgen.0.000294
18. Zhang T, Li H, Ma S *et al.* The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. *Appl Environ Microbiol* 2023;89:e0060523. https://doi.org/10.1128/aem.00605-23
19. Maguire F, Jia B, Gray KL *et al.* Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb Genomics* 2020;6:mgen000436. https://doi.org/10.1099/mgen.0.000436
20. Amarasinghe SL, Su S, Dong X *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020; 21:30. https://doi.org/10.1186/s13059-020-1935-5
21. Kolesnikov A, Cook D, Nattestad M *et al.* Local read haplotagging enables accurate long-read small variant calling. *Nat Commun* 2024;15:5907. https://doi.org/10.1038/s41467-024-50079-5
22. Method of the Year 2022: long-read sequencing. *Nat Methods* 2023;20:1. https://doi.org/10.1038/s41592-022-01759-x
23. Marić J, Križanović K, Riondet S *et al.* Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *BMC Bioinf* 2024;25:15. https://doi.org/10.1186/s12859-024-05634-8
24. 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
25. Kim C, Pongpanich M, Porntaveetus T Unraveling metagenomics through long-read sequencing: a comprehensive review. *J Transl Med* 2024;22:111. https://doi.org/10.1186/s12967-024-04917-1
26. Brown SD, Dreolini L, Wilson JF *et al.* Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinf* 2023;24:116. https://doi.org/10.1186/s12859-023-05226-y
27. Chakrawarti A, Eckstrom K, Laaguiby P *et al.* Hybrid Illumina-Nanopore assembly improves identification of multilocus sequence types and antimicrobial resistance genes of Staphylococcus aureus isolated from Vermont dairy farms: comparison to Illumina-only and R9.4.1 nanopore-only assemblies. *Access Microbiol* 2024; 6:000766.v3. https://doi.org/10.1099/acmi.0.000766.v3
28. Bertrand D, Shaw J, Kalathiyappan M *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* 2019;37:937–44. https://doi.org/10.1038/s41587-019-0191-2
29. Xie H, Yang C, Sun Y *et al.* PacBio long reads improve metagenomic assemblies, gene catalogs, and genome binning. *Front Genet* 2020;11:516269. https://doi.org/10.3389/fgene.2020.516269
30. Van Goethem MW, Osborn AR, Bowen BP *et al.* Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. *Commun Biol* 2021;4:1302. https://doi.org/10.1038/s42003-021-02809-4
31. Dai D, Brown C, Bürgmann H *et al.* Long-read metagenomic sequencing reveals shifts in associations of antibiotic resistance genes with mobile genetic elements from sewage to activated sludge. *Microbiome* 2022;10:20. https://doi.org/10.1186/s40168-021-01216-5
32. Bickhart DM, Kolmogorov M, Tseng E *et al.* Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat Biotechnol* 2022;40:711–9. https://doi.org/10.1038/s41587-021-01130-z
33. Wilbanks EG, Doré H, Ashby MH *et al.* Metagenomic methylation patterns resolve bacterial genomes of unusual size and

structural complexity. *ISME J* 2022; 16:1921–31. https://doi.org/10.1038/s41396-022-01242-7
34. Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids Res* 2024; 52:D777–83. https://doi.org/10.1093/nar/gkad943
35. Dmitrijeva M, Ruscheweyh H-J, Feer L *et al.* The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res* 2025; 53:D797–805. https://doi.org/10.1093/nar/gkae1004
36. Richardson L, Allen B, Baldi G *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023; 51:D753–9. https://doi.org/10.1093/nar/gkac1080
37. Chen I-MA, Chu K, Palaniappan K *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* 2023;51:D723–32. https://doi.org/10.1093/nar/gkac976
38. Rajwani R, Ohlemacher SI, Zhao G *et al.* Genome-guided discovery of natural products through multiplexed low-coverage whole-genome sequencing of soil actinomycetes on Oxford Nanopore Flongle. *Msystems* 2021;6:e01020–21. https://doi.org/10.1128/mSystems.01020-21
39. Sánchez-Navarro R, Nuhamunada M, Mohite OS *et al.* Long-read metagenome-assembled genomes improve identification of novel complete biosynthetic gene clusters in a complex microbial activated sludge ecosystem. *Msystems* 2022;7:e00632–22. https://doi.org/10.1128/msystems.00632-22
40. Roberts LW, Enoch DA, Khokhar F *et al.* Long-read sequencing reveals genomic diversity and associated plasmid movement of carbapenemase-producing bacteria in a UK hospital over 6 years. *Microb Genomics* 2023;9:mgen001048. https://doi.org/10.1099/mgen.0.001048
41. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res*. 2025;53:D20–9. https://doi.org/10.1093/nar/gkae979
42. O'Cathail C, Ahamed A, Burgin J *et al.* The European Nucleotide Archive in 2024. *Nucleic Acids Res* 2025; 53:D49–55. https://doi.org/10.1093/nar/gkae975
43. CNCB-NGDC Members and Partners . Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. *Nucleic Acids Res* 2025; 53:D30–44. https://doi.org/10.1093/nar/gkae978
44. Rimmer CO, Thomas JC. Complete genome sequence of *Staphylococcus edaphicus* Strain CCM 8731. *Microbiol Resour Announc* 2022;11:e0051822. https://doi.org/10.1128/mra.00518-22
45. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinforma Oxf Engl* 2023;39:btad311. https://doi.org/10.1093/bioinformatics/btad311
46. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinforma Oxf Engl* 2021;37:4572–4. https://doi.org/10.1093/bioinformatics/btab705
47. Che.n S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* 2023;2:e107. https://doi.org/10.1002/imt2.107
48. Langmead B, Wilks C, Antonescu V *et al.* Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinforma Oxf Engl* 2019;35:421–32. https://doi.org/10.1093/bioinformatics/bty648
49. Shen W, Sipos B, Zhao L. SeqKit2: a Swiss army knife for sequence and alignment processing. *Imeta* 2024;3:e191. https://doi.org/10.1002/imt2.191
50. Lu J, Rincon N, Wood DE *et al.* Metagenome analysis using the Kraken software suite. *Nat Protoc* 2022;17:2815–39. https://doi.org/10.1038/s41596-022-00738-y
51. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinf* 2011;12:385. https://doi.org/10.1186/1471-2105-12-385
52. Feldgarden M, Brover V, Gonzalez-Escalona N *et al.* AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* 2021;11:12728. https://doi.org/10.1038/s41598-021-91456-0
53. Zhou S, Liu B, Zheng D *et al.* VFDB 2025: an integrated resource for exploring anti-virulence compounds. *Nucleic Acids Res* 2025; 53:D871–7. https://doi.org/10.1093/nar/gkae968
54. Peng K, Gao Y, Li C *et al.* Benchmarking of analysis tools and pipeline development for nanopore long-read metagenomics. *Sci Bulletin* 2025;70:1591–5. https://doi.org/10.1016/j.scib.2025.03.044
55. Nurk S, Meleshko D, Korobeynikov A *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017; 27:824–34. https://doi.org/10.1101/gr.213959.116
56. Kolmogorov M, Bickhart DM, Behsaz B *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;17:1103–10. https://doi.org/10.1038/s41592-020-00971-x
57. Hu J, Fan J, Sun Z *et al.* NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinforma Oxf Engl* 2020;36:2253–5. https://doi.org/10.1093/bioinformatics/btz891
58. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinforma Oxf Engl* 2016;32:1088–90. https://doi.org/10.1093/bioinformatics/btv697
59. Hyatt D, LoCascio PF, Hauser LJ *et al.* Gene and translation initiation site prediction in metagenomic sequences. *Bioinforma Oxf Engl* 2012;28:2223–30. https://doi.org/10.1093/bioinformatics/bts429
60. Fu L, Niu B, Zhu Z *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma Oxf Engl* 2012;28:3150–2. https://doi.org/10.1093/bioinformatics/bts565
61. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;18:366–8. https://doi.org/10.1038/s41592-021-01101-x
62. Kahlke T, Ralph PJ. BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol Evol* 2019;10:100–3. https://doi.org/10.1111/2041-210X.13095
63. Patro R, Duggal G, Love MI *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14:417–9. https://doi.org/10.1038/nmeth.4197
64. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293
65. Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; 47:D309–14. https://doi.org/10.1093/nar/gky1085
66. Drula E, Garron M-L, Dogan S *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* 2022; 50:D571–7. https://doi.org/10.1093/nar/gkab1045
67. Alcock BP, Huynh W, Chalil R *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2023; 51:D690–9. https://doi.org/10.1093/nar/gkac920
68. Bai D, Chen T, Xun J *et al.* EasyMetagenome: a user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. *Imeta* 2025;4:e70001. https://doi.org/10.1002/imt2.70001
69. Pan S, Zhao X-M, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinforma Oxf Engl* 2023;39:i21–9. https://doi.org/10.1093/bioinformatics/btad209
70. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018;6:158. https://doi.org/10.1186/s40168-018-0541-1

71. Sieber CMK, Probst AJ, Sharrar A *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;3:836–43. https://doi.org/10.1038/s41564-018-0171-1
72. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;20:1203–12. https://doi.org/10.1038/s41592-023-01940-w
73. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinforma Oxf Engl* 2022;38:5315–6. https://doi.org/10.1093/bioinformatics/btac672
74. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl* 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153
75. Stothard P, Grant JR, Van Domselaar G. Visualizing and comparing circular genomes using the CGView family of tools. *Brief Bioinform* 2019;20:1576–82. https://doi.org/10.1093/bib/bbx081
76. Gong J, Zhang Q, Hu R *et al.* Effects of *Prevotella copri* on insulin, gut microbiota and bile acids. *Gut Microbes* 2024;16:2340487. https://doi.org/10.1080/19490976.2024.2340487
77. Chen C, Fang S, Wei H *et al. Prevotella copri* increases fat accumulation in pigs fed with formula diets. *Microbiome* 2021;9:175. https://doi.org/10.1186/s40168-021-01110-0
78. Abdelsalam NA, Hegazy SM, Aziz RK. The curious case of *Prevotella copri*. *Gut Microbes* 2023;15:2249152. https://doi.org/10.1080/19490976.2023.2249152
79. Castillo Y, Delgadillo NA, Neuta Y *et al.* Antibiotic susceptibility and resistance genes in oral clinical isolates of *Prevotella intermedia*, *Prevotella nigrescens*, and *Prevotella melaninogenica*. *Antibiot Basel Switz* 2022;11:888. https://doi.org/10.3390/antibiotics11070888
80. Tuson HH, Foley MH, Koropatkin NM *et al.* The starch utilization system assembles around stationary starch-binding proteins. *Biophys J* 2018;115:242–50. https://doi.org/10.1016/j.bpj.2017.12.015
81. Benz JP, Chau BH, Zheng D *et al.* A comparative systems analysis of polysaccharide-elicited responses in Neurospora crassa reveals carbon source-specific cellular adaptations. *Mol Microbiol* 2014;91:275–99. https://doi.org/10.1111/mmi.12459
82. Li J, Gálvez EJC, Amend L *et al.* A versatile genetic toolbox for *Prevotella copri* enables studying polysaccharide utilization systems. *EMBO J* 2021; 40:e108287. https://doi.org/10.15252/embj.2021108287
83. Scott KP, Gratz SW, Sheridan PO *et al.* The influence of diet on the gut microbiota. *Pharmacol Res* 2013;69:52–60. https://doi.org/10.1016/j.phrs.2012.10.020
84. Kim KN, Yao Y, Ju SY Short chain fatty acids and fecal microbiota abundance in humans with obesity: a systematic review and meta-analysis. *Nutrients* 2019;11:2512. https://doi.org/10.3390/nu11102512
85. Berger B, Haas D Transposase and cointegrase: specialized transposition proteins of the bacterial insertion sequence IS21 and related elements. *Cell Mol Life Sci.* 2001;58:403–19. https://doi.org/10.1007/PL00000866
86. Kirsch JM, Hryckowian AJ, Duerkop BA A metagenomics pipeline reveals insertion sequence-driven evolution of the microbiota. *Cell Host Microbe* 2024;32:739–54. https://doi.org/10.1016/j.chom.2024.03.005
87. Almeida A, Nayfach S, Boland M *et al.* A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;39:105–14. https://doi.org/10.1038/s41587-020-0603-3
88. Kim CY, Lee M, Yang S *et al.* Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med* 2021; 13:134. https://doi.org/10.1186/s13073-021-00950-7
89. Udaondo Z, Sittikankaew K, Uengwetwanit T *et al.* Comparative analysis of PacBio and Oxford Nanopore Sequencing Technologies for transcriptomic landscape identification of Penaeus monodon. *Life* 2021;11:862. https://doi.org/10.3390/life11080862
90. Tafazoli A, Hemmati M, Rafigh M *et al.* Leveraging long-read sequencing technologies for pharmacogenomic testing: applications, analytical strategies, challenges, and future perspectives. *Front Genet* 2025;16:1435416. https://doi.org/10.3389/fgene.2025.1435416
91. Kang X, Xu J, Luo X *et al.* Hybrid-hybrid correction of errors in long reads with HERO. *Genome Biol* 2023; 24:275. https://doi.org/10.1186/s13059-023-03112-7
92. Latorre-Pérez A, Villalba-Bermell P, Pascual J *et al.* Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci Rep* 2020;10:13588. https://doi.org/10.1038/s41598-020-70491-3
93. Logsdon GA, Vollger MR, Eichler EE Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;21:597–614. https://doi.org/10.1038/s41576-020-0236-x
94. Gand M, Bloemen B, Vanneste K *et al.* Comparison of 6 DNA extraction methods for isolation of high yield of high molecular weight DNA suitable for shotgun metagenomics Nanopore sequencing to detect bacteria. *Bmc Genomics [Electronic Resource]* 2023;24:438. https://doi.org/10.1186/s12864-023-09537-5
95. De La Cerda GY, Landis JB, Eifler E *et al.* Balancing read length and sequencing depth: optimizing Nanopore long-read sequencing for monocots with an emphasis on the Liliales. *Appl Plant Sci* 2023;11:e11524. https://doi.org/10.1002/aps3.11524
96. Sauvage T, Cormier A, Delphine P A comparison of Oxford nanopore library strategies for bacterial genomics. *Bmc Genomics [Electronic Resource]* 2023;24:627. https://doi.org/10.1186/s12864-023-09729-z
97. Sanderson ND, Kapel N, Rodger G *et al.* Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genomics* 2023;9:mgen000910. https://doi.org/10.1099/mgen.0.000910

Received: August 11, 2025. **Revised:** October 10, 2025. **Accepted:** October 11, 2025