

MetaflowX: a scalable and resource-efficient workflow for multi-strategy metagenomic analysis

Yan Xia^{1,2,3,†}, Lifeng Liang^{1,2,†}, Xiaokai Wang^{1,2,†}, Zixiang Chen^{1,†}, Jin Liu⁴, Ying Yang¹, Hailiang Xie¹, Zhimin Ding¹, Xiaoting Huang¹, Shibin Long¹, Zhifeng Wang¹, Xiaoqiang Xu¹, Chao Ding^{5,*}, Qiyi Chen^{6,7,8,*}, Qiang Feng^{9,*}

¹01Life Institute, Shenzhen 518100, China

²Shen 1001 Life Institute, Shanghai 200435, China

³State Key Laboratory of Pharmaceutical Biotechnology, Chemistry and Biomedicine Innovation Center (ChemBIC), School of Life Sciences, Nanjing University, Nanjing 210093, China

⁴Department of Life Sciences, Yuncheng University, Yuncheng, Shanxi 044000, China

⁵Department of General Surgery, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, China

⁶Department of Functional Intestinal Diseases, Department of General Surgery, Shanghai, Tenth People's Hospital, Tongji University School of Medicine, Shanghai 200072, China

⁷Shanghai Gastrointestinal Microecology Research Center, Shanghai 200072, China

⁸Shanghai Institution of Gut Microbiota Research and Engineering Development, Shanghai 200435, China

⁹Department of Human Microbiome, School and Hospital of Stomatology, Cheeloo College of Medicine, Shandong University & Shandong Key Laboratory of Oral Tissue Regeneration & Shandong Engineering Research Center of Dental Materials and Oral Tissue Regeneration & Shandong Provincial Clinical Research Center for Oral Diseases, Shandong 250012, China

*To whom correspondence should be addressed. Email: dingchao21@nju.edu.cn

Correspondence may also be addressed to Qiyi Chen. Email: qiyichen2011@163.com

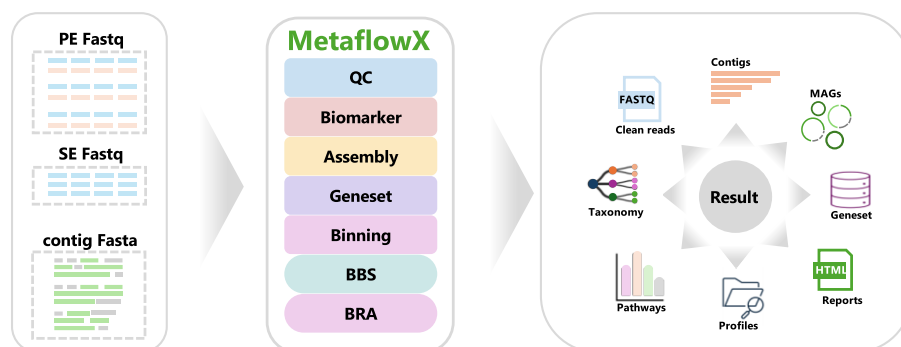
Correspondence may also be addressed to Qiang Feng. Email: fengqiangsdu@163.com

[†]These authors contributed equally to this work.

Abstract

Microbiomes play crucial roles in diverse ecosystems, spanning environmental, agricultural, and human health domains. However, in-depth metagenomic data analysis presents significant technical and resource challenges, particularly at scale. Existing computational pipelines are typically limited to either reference-based or reference-free approaches and exhibit inefficiencies in process large datasets. Here, we introduce MetaflowX (<https://github.com/01life/MetaflowX>), an open-resource workflow integrating both analytical paradigms for enhanced metagenomic investigations. This modular framework encompasses short-read quality control, rapid microbial profiling, hybrid contig assembly and binning, high-quality metagenome-assembled genome (MAG) identification, as well as bin refinement and reassembly. Benchmarking tests showed that MetaflowX completed full metagenomic analyses up to 14-fold faster and with 38% less disk usage than existing workflows. It also recovered the highest number of high-quality and taxonomically diverse MAGs. A dedicated reassembly module further improved MAG quality, increasing completeness by 5.6% and reducing contamination by 53% on average. Functional annotation modules enable detection of key features, including virulence and antibiotic resistance genes. Designed for extensibility, MetaflowX provides an efficient solution addressing current and emerging demands in large-scale metagenomic research.

Graphical abstract



Received: November 12, 2024. Accepted: August 23, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

Microorganisms are ubiquitous in nature and the human body, playing a pivotal role in maintaining environmental equilibrium and human health [1–8]. Next-generation sequencing (NGS) technology provides a culture-independent approach for investigating microbial taxonomic composition and potential metabolic functions [9]. Among these applications, comprehensive analysis of metagenomics data constitutes a crucial yet time-intensive phase. Over the past decades, numerous computational methodologies have been rapidly developed, including microbial profiling, contig assembly, contig binning, gene prediction, and function assigning [10–12], which have substantially advanced our ability to analyze and interpret metagenomic datasets.

Taxonomic characterization constitutes a crucial step in microbiome data analysis. Numerous computational pipelines have been developed for taxonomic classification and metagenome-assembled genomes (MAGs) retrieval, including metaWRAP [13], bioBakery [14, 15], ATLAS [16], nf-core/taxprofiler [17], MAGNETO [18], and nf-core/mag(nf-mag) [19]. Among these, MetaPhlAn4 [20], HUMAnN3 [15] (components of bioBakery), and nf-core/taxprofiler employ reference-based methodologies that rely on aligning sequencing reads to isolate genomes, gene catalogs, or established reference marker gene datasets. In contrast, metaWRAP, ATLAS, MAGNETO, and nf-mag, utilize reference-free approaches. Reference-free pipelines typically involve preprocessing steps such as contig assembly and binning to reconstructed MAGs, followed by taxonomic and functional annotation. Furthermore, reference-free pipeline like MUFFIN [21], ATLAS, nf-mag utilize workflow management systems such as Snake-make [22] or Nextflow [23] to improve scalability, portability, reproducibility, and user-friendliness. These tools have proven particularly effective for studying uncultivated microbial communities. Conversely, in well-characterized environments with comprehensive reference databases, such as human and murine gut microbiomes, reference-based methods often suffice to extract biological insights while avoiding the computational overhead associated with MAGs reconstruction.

Metagenomic assembly rarely produces complete single-contig genomes for microbial taxa, necessitating contig clustering through binning procedures to reconstruct MAGs. Contemporary binning algorithms predominantly utilize sequence-intrinsic features including k -mer composition and coverage profiles. Established tools such as MetaBAT2 [24], MaxBin2 [25], MetaBinner [26], and MetaDecoder [27] employ probabilistic models or graph-based models to process these features. Emerging methodologies like SemiBin [28, 29], VAMB [30], and COMEBin [31] leverage deep neural networks to extract higher-order patterns from sequencing data [32]. Integrative frameworks such as DAS Tool [33], MAGScoT [34], and the “bin-refinement” module in metaWRAP, amalgamate findings from various binning techniques to improve precision. Deepurify [35], MetaCC [36], and BASALT [37] provide additional enhancements in binning precision. However, a versatile workflow that effectively incorporates multiple binners, reassembles, and refines MAGs is still lacking. To address this critical gap, we introduced a workflow named MetaflowX (Fig. 1), which integrates both reference-based and reference-free methods, along with workflow managers, to better address the current needs of metagenomic research.

Material and methods

MetaflowX architecture

MetaflowX is a Nextflow-based metagenomics analysis workflow that processes short-read sequences and/or assembled contigs to automatically generate taxonomic compositions, community functional profiles, non-redundant gene catalogs with functional annotations, and high-quality (HQ) MAGs [38]. The workflow consists of five key modules as depicted in Fig. 1 and Supplementary Fig. S1. Module 1 (MetaflowX-QC module) performs quality control on raw sequence data using established tools such as fastp [39] and Trimmomatic [40], with the aim of excluding low-quality (LQ) reads and potential contaminant or host sequences (Supplementary Fig. S2A). Module 2 (MetaflowX-Assembly module) utilizes metaSPAdes [41] and MEGAHIT [42] to assemble reads into contigs (Supplementary Fig. S2C). Contigs longer than 2000 bp are processed by Modules 4 and 5 (MetaflowX-Geneset and MetaflowX-Binning module). Module 3 (MetaflowX-BioMarker module) employs a reference-based approach to generate taxonomic, gene, and pathway abundances, integrating taxonomic profilers like MetaPhlAn4 and Kraken2 [43, 44], along with the functional profiler HUMAnN3 (Supplementary Fig. S2B). Module 4 (MetaflowX-Geneset module) generates a non-redundant gene set for the microbial community, leveraging diverse orthology databases (eggNOG [45], COG [46, 47], KEGG [48], CAZy [49], CARD [50], and VFDB [51]) to assign functions and estimate gene abundance (Supplementary Fig. S2D). Module 5 (Binning module) generates genome assemblies for individual microorganisms, integrating results from multiple binners to recover HQ MAGs and profile their taxonomy (Supplementary Figs S3 and S4). A more detailed description was provided in Supplementary Results, section “Details of the MetaflowX implementation and workflow.”

To ensure MetaflowX’s accessibility and reliability across different scenarios, we designed comprehensive evaluation strategies spanning various computing environments, from local workstations to high-performance computing clusters and major cloud platforms (including Amazon Web Services (AWS), Alibaba Cloud, and Tencent Cloud), coupled with thorough negative testing protocols.

Implementation

Installation and configuration

MetaflowX integrates over 30 bioinformatic tools to enable comprehensive analyses (Supplementary Table S2). Given the complex interdependencies among these tools, creating a single operational environment is unfeasible. To address this challenge, MetaflowX adopts a modular approach, compartmentalizing the tools into independent Conda environments that require approximately 6 GB of storage space. In addition to the software suite, MetaflowX utilizes 15 reference databases essential for taxonomic identification and functional annotation (Supplementary Table S3). These databases require an additional 436.6 GB of storage for pre-construction. We strongly recommend that users pre-build both the software components and reference databases before running MetaflowX analyses. This approach differs from that of other pipelines, such as nf-mag, which opt for dynamic download during runtime. Pre-construction is crucial for minimizing setup time and

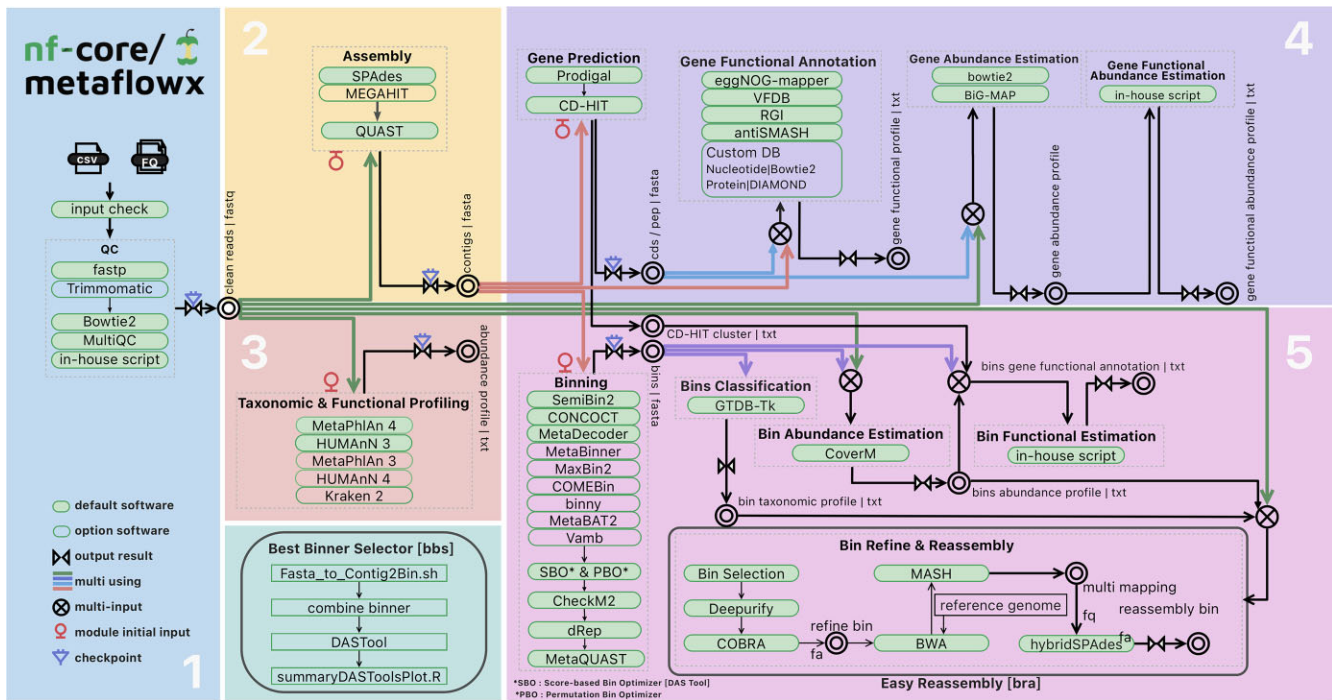


Figure 1. Schematic overview of the metaflowX workflow. The MetaflowX workflow consists of a standard pipeline with five key modules as Panels 1–5: (1) quality control (blue), (2) contig assembly (yellow), (3) microbial taxonomy and metabolic function analysis (red), (4) gene catalog construction (purple), and (5) automated binning analysis (pink). Additionally, the workflow includes two mini-tools: Best Binner Detector (green, rounded rectangle below Panel 4) and Bin Refine and Reassembly (pink, rounded rectangle within Panel 5). The default software associated with each module is depicted by fully rounded green rectangles, while optional software is represented by rounded green rectangles. Data flows that are utilized multiple times are illustrated by colorful, strong lines. Specific markers are used to denote each module's initial input, output, and checkpoints.

improving the efficiency of bioinformatic analyses. Detailed guidelines for software installation and database construction are available in the MetaflowX documentation (<https://github.com/01life/MetaflowX/blob/main/docs/dependencies.md> and <https://github.com/01life/MetaflowX/blob/main/docs/database.md>).

Advanced error handling for robust operation

MetaflowX incorporates a pre-execution check mechanism designed to identify and prevent configuration errors and data format inconsistencies before job execution. These checks validate input and software configurations, enforce sample ID naming conventions, verify the integrity of gz-compressed files, and assess the balance of paired-end read volumes—thereby reducing runtime error rates and enhancing pipeline robustness. Failures in processes such as read quality control, contig assembly, MetaPhlAn, and HUMAnN will trigger pipeline halts, allowing for necessary manual intervention. However, the binning process only generates warning messages and does not affect other analyses.

MetaflowX incorporates strategies to manage storage resource constraints and I/O bottlenecks. On cloud computing platforms, it includes safeguards against instance disconnections caused by service failures. The pipeline adopts a default “errorStrategy” of “retry” upon initial failure, dynamically adjusting CPU and memory resources based on “exitcode” assessments. The final retry strategy is set to “finish,” enabling all tasks to complete before termination and thereby isolating erroneous tasks to prevent them from impacting parallel operations. Errors are documented in “MetaflowX_error/warning_log.txt,” providing

users with guidance in addressing failed samples or tasks (Supplementary Fig. S5).

Integration of multiple binning algorithms for mag reconstruction

To reduce algorithm-specific biases and recover a more comprehensive set of MAGs, MetaflowX incorporates two bin refinement strategies: Scored Bin Optimizer (SBO) and Permutation Bin Optimizer (PBO). These strategies enable flexible integration of outputs from multiple binning algorithms within the Module 5 (Binning module) of MetaflowX. Currently, MetaflowX supports a wide range of binners, including MetaBAT2, MaxBin2, CONCOCT, MetaBinner, binny, COMEBin, SemiBin2, MetaDecoder, and Vamb. Among these, MetaDecoder, CONCOCT, and SemiBin2 are set as default tools based on both internal benchmarking presented in Supplementary Results (section “Evaluation of MetaflowX-Binning performance for MAG recovery,” Supplementary Tables S10–S12, and Supplementary Figs S9–S15) and previous studies [52].

The SBO approach relies on quality-based selection, utilizing either single-copy marker genes or k-mer-based estimators to assess completeness and contamination. It leverages existing tools such as DAS Tool or MAGScoT for scoring and dereplicating bins, ultimately generating a non-redundant set of HQ MAGs.

The PBO approach, by contrast, is a marker gene-free strategy that integrates binning results from multiple tools via a path-based intersection method. For each sample, contigs are first sorted by length and then processed sequentially. For each contig, PBO retrieves the clusters assigned by each tool, col-

lects all contigs within those clusters, and calculates their intersection. If the intersection is non-empty, the overlapping contigs are assigned to a consensus bin. Contigs failing to meet the consistency criteria are excluded through a non-replacement filtering process, ensuring only high-confidence, non-redundant assignments are retained. Bins with a total sequence length of <0.5 Mbp or >20 Mbp are discarded to reduce incomplete or potentially chimeric assemblies. The remaining bins are evaluated using CheckM2, with only those exhibiting an estimated completeness > 50% and contamination < 10% being retained.

When both the SBO and PBO strategies are enabled, MetaflowX aggregates all HQ MAGs identified by either method, with selection based on CheckM2 evaluation (completeness > 50% and contamination < 10%). A representative MAG is then selected from each cluster using Galah (<https://github.com/wwood/galah>), which prioritizes genomes according to user-defined criteria.

Benchmarking of metaflowx and comparative workflows

To evaluate the performance of MetaflowX, we conducted two benchmarking analyses using short-read datasets from the Critical Assessment of Metagenome Interpretation (CAMI) II initiative (<https://www.cami-challenge.org>).

First, for a comprehensive end-to-end workflow comparison, ten human gut metagenomes from the CAMI II Toy Human Microbiome Project (HMP) dataset were processed using four complete workflows: MetaflowX, MetaWRAP, nf-mag, and anvi'o [53, 54]. All workflows were executed with their respective default parameters to ensure a fair comparison of the entire analytical pipeline.

Second, to assess binning performance across diverse simulated metagenomic contexts, we performed targeted comparisons using three workflows: MetaflowX, MetaWRAP, and nf-mag. Four CAMI II datasets were selected: Plant-associated (Plant), Marine, Toy Mouse Gut (MG), and Toy Human Microbiome Project (HMP). From each dataset, ten samples were randomly selected, yielding a total of 40 samples for binning evaluation. To ensure consistency and compatibility across binning algorithms, only assembled contigs longer than 2000 bp were retained for downstream analysis. MAG quality assessment was performed using CheckM2 for all genomes generated by MetaflowX, MetaWRAP, and nf-mag. Genome quality scores (QS) were calculated as completeness – 5 × contamination, following established practice [4]. To assess redundancy and clustering resolution, MAGs were dereplicated at both the species level (Average Nucleotide Identity (ANI) ≥ 95%) and strain level (ANI ≥ 99%) using Galah. Species-level dereplication was performed with the parameters “–ani 95 –precluster-ani 90,” and strain-level with “–ani 99 –precluster-ani 95.” Taxonomic classification of dereplicated MAGs was conducted using GTDB-Tk against the Genome Taxonomy Database (GTDB) release r226.

Real datasets for completeness of genome re-recovery assessment

To evaluate the efficacy of RefineReassembly (DR) in recovering bin genomes, 23 real metagenomic datasets from the Human Early Life Cohort project were used. Raw sequencing data were obtained from the European Nucleotide Archive (ENA) using provided accession numbers

(Supplementary Table S4). The data underwent preprocessing, assembly, and binning via MetaflowX prior to being processed by the DR module. Due to the large volume of sequencing data, MEGAHIT was used as the assembler instead of the default SPAdes. Contigs longer than 2 000 base pairs were retained, resulting in a total of 581 588 contigs. Initial binning with default software generated 428 bins, with their specific taxonomic classifications and quality information detailed in Supplementary Table S5. The DR pipeline was running using default parameters (–minCompleteness 90 –minContamination 5 –minQS 65 –minCount 10 000 –minDepth 1), selecting 110 bins for co-assembly. For the single-sample assembly step, modified parameters were applied: get_bin_assembly_options = “–minCompleteness 90 –minContamination 5 –minQS 65 –minCount 10 000 –minDepth 1 –singleAssembly” and extract_bin_reads_options = “–use_single_sample –max_dist_threshold 0.2 –topSampleNum 10.” Quality evaluations at each process stage were performed using CheckM2, assessing bin completeness and contamination scores. The progression of bin quality throughout the DR module was visualized using a Sankey diagram generated with the R (version 4.3.2) package ggsankey.

Identification of ARGs and VFGs

We identified antibiotic resistance genes (ARGs) and virulence factor genes (VFGs) in 28 metagenomic samples derived from the intestinal fluid of seven patients with intestinal obstruction (unpublished data). Specifically, we used the Resistance Gene Identifier (RGI, version 6.0.2) with default settings to detect ARGs within the bins, leveraging the Comprehensive Antibiotic Resistance Database (CARD, version 3.2.7). VFGs were identified using DIAMOND blastp (version 2.0.15) against the VFDB database. ARG and VFG categories were assigned based on the sequence descriptions in the respective datasets.

Statistics and visualization

All statistical analyses were performed using R (version 4.3.2). Complex visualizations for the article were generated utilizing the ggplot2 (version 3.5.1) and ggpubr (version 0.6.0) packages. For differential analyses, Wilcoxon signed-rank tests were applied to paired data and Wilcoxon rank-sum tests to unpaired data, with both implemented via the “stat_compare_means” function in ggplot2.

Within the MetaflowX pipeline, data summaries and intermediate statistics were generated using pandas (version 2.0.3) in Python (version 3.10.11). Interactive visualizations included in the HTML reports were created with Plotly (version 5.15.0) for complex, interactive charts, and Bokeh (version 3.3.4) for customizable, web-based plots.

Results

MetaflowX is a fast and flexible workflow optimized for large-scale metagenomic data analysis

To address the computational and organizational challenges in large-scale metagenomic studies, we developed MetaflowX—a fast and flexible workflow built on the Nextflow framework. It integrates 37 tools across 142 tasks, which are organized into 18 subworkflows under five functional modules: quality control, assembly, gene catalog construction, binning, and taxonomic or functional profiling. This

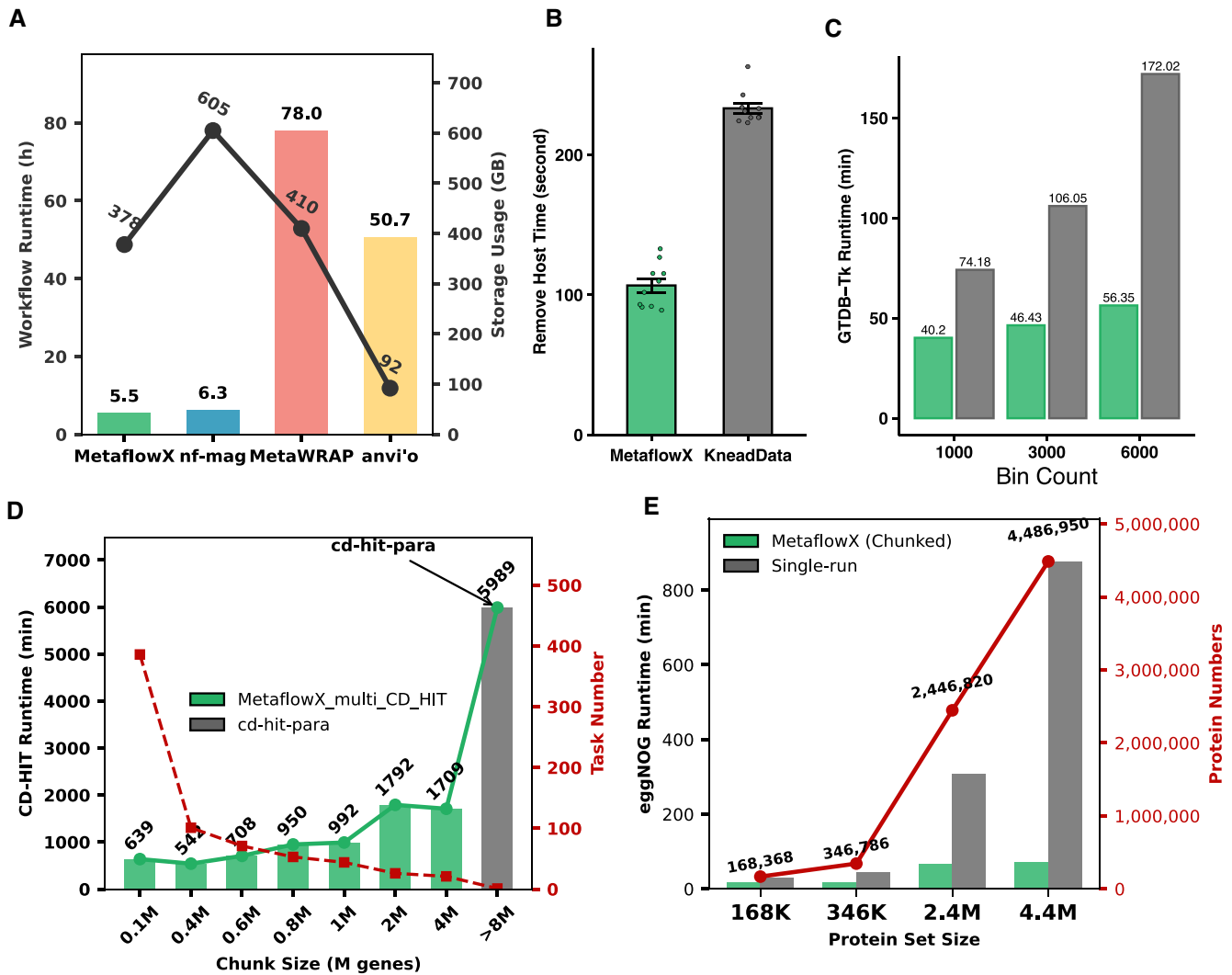


Figure 2. MetaflowX achieves improved runtime and reduced resource usage through algorithmic improvements and modular workflow optimization. (A) Comparison of total runtime (bars) and peak storage usage (line) among four metagenomic workflows—MetaflowX, nf-mag, MetaWRAP, and anvi'o—on 10 CAMI II HMP samples. (B) Host read removal runtime comparison between MetaflowX and KneadData, based on alignment of reads to the host genome using Bowtie2. (C) Runtime comparison of GTDB-Tk taxonomy annotation with and without bin set chunking. MetaflowX splits bin sets into chunks of 500 for parallel execution, reducing annotation time across varying input sizes (1000–6000 bins). (D) Total runtime for clustering gene catalogs using CD-HIT. MetaflowX performs chunked CD-HIT clustering with different chunk sizes (0.1–8 M genes), showing improved efficiency compared to monolithic cd-hit-para. The number of tasks required for each chunk size is indicated by the dashed line (right axis). (E) Runtime comparison of EggNOG-mapper functional annotation for protein sequences using chunked versus single-run strategies.

modular design supports over 30 configurable usage scenarios (Supplementary Fig. S6), including targeted workflows such as non-redundant gene set construction, rapidly updating bin taxonomy identification, and estimating gene and bin abundances via alternative methods.

In our full-workflow benchmark using ten CAMI II HMP samples (Fig. 2A), MetaflowX achieved the shortest total runtime (5.5 h), compared to 6.3 h for nf-mag, 50.7 h for anvi'o, and 78.0 h for MetaWRAP. This corresponds to runtime reductions of 89.2% and 93.0% compared to anvi'o and MetaWRAP, respectively. These performance gains stem largely from the task orchestration capabilities of the Nextflow engine, which enables efficient parallelization and minimizes idle time, in contrast to the *ad hoc* scheduling in MetaWRAP or the Snakemake-based design of anvi'o. In terms of disk usage, MetaflowX consumed 378 GB, representing a 37.5% reduction compared to nf-mag (605 GB), and be-

ing comparable to MetaWRAP (410 GB). This efficiency is attributable to the automatic cleanup of intermediate files after each subtask, a feature absent in nf-mag. Although anvi'o exhibited the lowest overall storage footprint (92 GB), this was due to its limited analytical scope and incomplete pipeline coverage. A comparison of MetaflowX's key features with those of other commonly used pipelines for shotgun metagenomic analysis is provided in Supplementary Table S1.

To maintain performance at scale, MetaflowX incorporates specific optimizations for high-complexity tasks. During host read removal, a custom Python script was developed to directly extract host-aligned reads from Bowtie2 output, reducing runtime by 54.3% compared to KneadData (Fig. 2B). For bin-level taxonomic identification using GTDB-Tk, large bin sets are partitioned into parallel jobs of 500 bins each; this reduced the annotation time for a 6000-bin dataset from 172.0 to 56.4 min, representing a 3-fold speedup (Fig. 2C). In gene

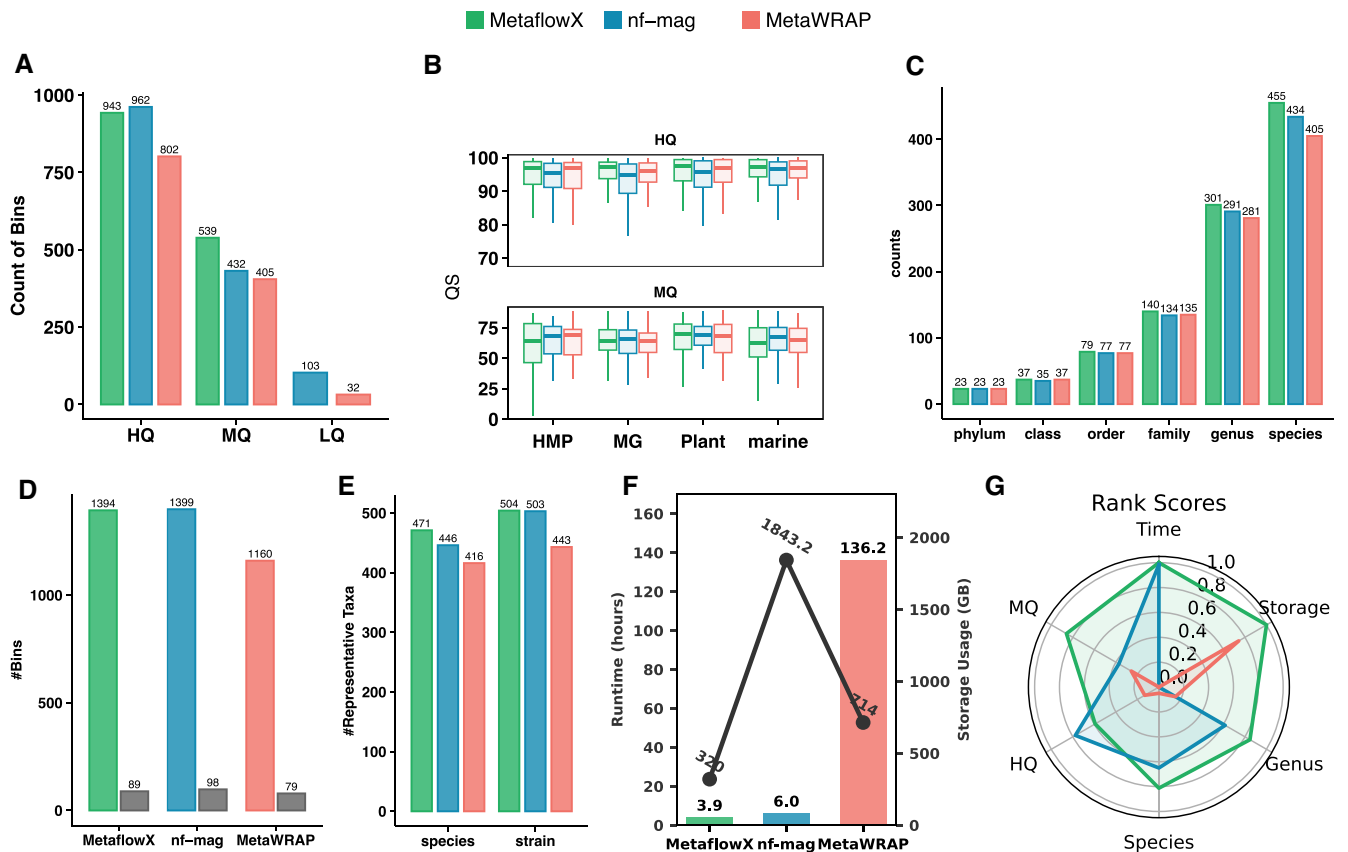


Figure 3. Evaluation of metaflowX binning performance compared to nf-mag and metawrap. **(A)** Total number of recovered bins across quality categories: HQ, MQ, and LQ, as assessed by CheckM2. **(B)** Distribution of quality scores (QS = completeness – 5 × contamination) for HQ and MQ bins across four CAMI II datasets: Human Microbiome Project (HMP), Mouse Gut (MG), Plant-associated, and Marine. **(C)** Number of taxonomically classified bins at different ranks (phylum to species) based on GTDB-Tk annotations. **(D)** Total number of bins classified at species level and unclassified by GTDB-Tk. **(E)** Number of non-redundant representative genomes at species (ANI ≥ 95%) and strain (ANI ≥ 99%) levels, determined using Galah. **(F)** Runtime (bars, in hours) and storage usage (line, in GB) for the binning and downstream processing modules across workflows. **(G)** Radar plot comparing the overall performance of MetaflowX, nf-mag, and MetaWRAP across six evaluation metrics: HQ bin count, MQ bin count, species recovery, genus recovery, runtime, and storage usage. All values were normalized to the range [0, 1] for comparison.

catalog clustering, we enhanced the parallelization scheme of the CD-HIT chunked execution tasks, thereby increasing task-level concurrency and overall computational efficiency. Consequently, chunked execution with a 1-million-gene chunk size reduced the runtime of cd-hit-para from 5989 to 992 min, corresponding to an 83.4% reduction (Fig. 2D). For protein function annotation with eggNOG-mapper, chunking offered minimal benefit for datasets with fewer than 350 000 proteins but yielded substantial improvements for larger datasets. Annotating 4.4 million proteins required only 67.5 min with chunking, versus 876.9 min in a single-task run—an impressive 92.3% reduction in runtime (Fig. 2E). Collectively, these results demonstrated that MetaflowX achieves enhanced computational efficiency through targeted algorithmic and workflow-level optimizations, making it well suited for high-throughput metagenomic analyses.

Evaluation of metaflowX-binning performance for MAG recovery

To evaluate the effectiveness of MetaflowX in reconstructing MAGs, we benchmarked its binning module against nf-mag and MetaWRAP using 40 metagenomic samples from four CAMI II datasets (HMP, MG, Plant, Marine). Anvi'o was ex-

cluded from this comparison due to its incomplete support for binning. All three workflows generated over 1000 bins: MetaflowX yielded 1483 total bins, a number comparable to nf-mag (1497) and exceeding MetaWRAP (1239) (Fig. 3A). However, there were differences in the distribution of bin quality, reflecting the distinct refinement strategies employed by each workflow. MetaflowX implements two complementary refinement strategies in parallel: the SBO, which selects bins based on QSs, and the PBO, which identifies consensus bins across multiple binners. Bins derived from both strategies are subsequently evaluated using CheckM2. MetaflowX recovered 943 HQ and 539 medium-quality (MQ) bins, with all LQ bins excluded. In contrast, nf-mag produced 103 LQ bins despite using DAS Tool refinement, attributable to the absence of post-hoc quality filtering. MetaWRAP integrates CheckM1-based filtering but still retained 32 LQ bins, likely due to algorithmic differences and threshold variations between CheckM1 and CheckM2 assessment frameworks. The multi-step bin selection strategy in MetaflowX minimizes false positives while enhancing genome quality through consensus-based validation. Across all environmental datasets, MetaflowX maintained a favorable QS distribution (Fig. 3B), particularly for MQ bins. Here, the stringent contig intersection in the PBO strategy slightly reduced completeness

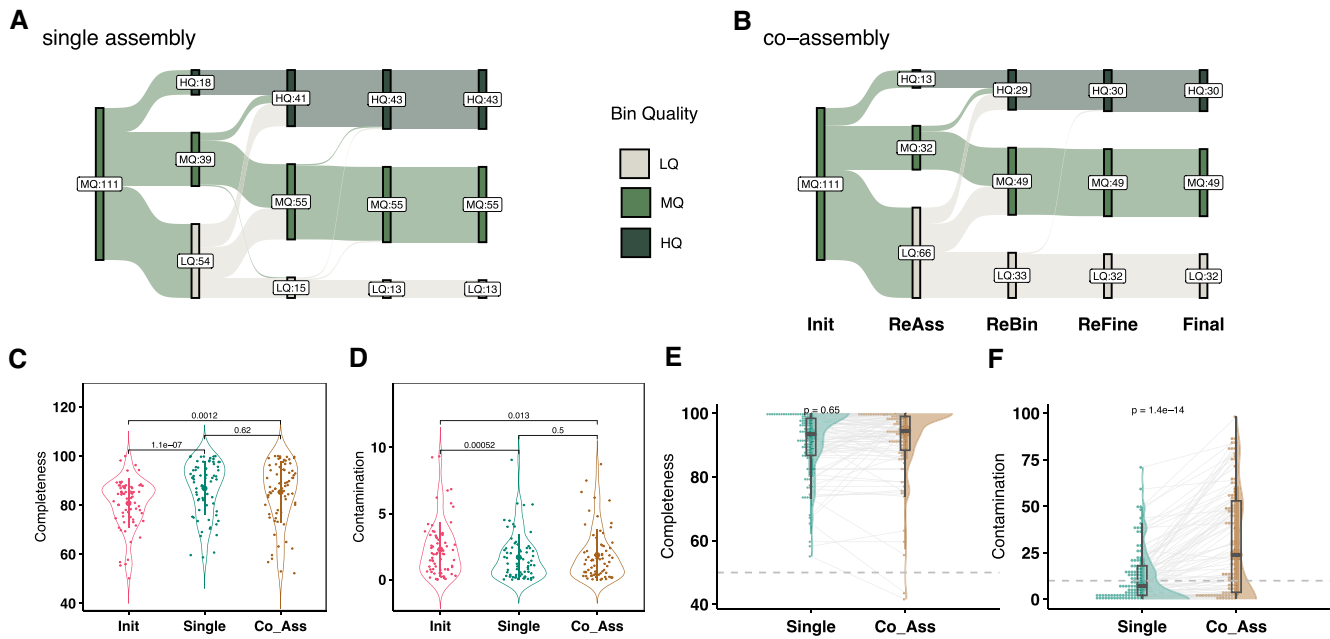


Figure 4. Comparative analysis of metaflowx-refinereassembly (DR) using co-assembly versus single assembly strategies for MAG recovery. Sankey diagrams showing bin quality changes across MetaflowX-DR stages using single assembly (**A**) and co-assembly (**B**) strategies. Stages include Initial binning (Init), ReAssembly (ReAss), ReBinning (ReBin), Refinement (ReFine), and Final binning. Bin quality categories: HQ, MQ, and low quality (LQ). Connection thickness represents bin count transitioning between stages. Violin plots showing completeness (**C**) and contamination (**D**) distributions for Initial binning (Init), single assembly (single), and co-assembly (Co_Ass) methods. Paired comparisons of completeness (**E**) and contamination (**F**) between single assembly and co-assembly. Composite graphs show differences between methods: dot plots (left), box plots (middle), and violin plots (right), with lines connecting paired bins.

but consistently minimized contamination (Supplementary Fig. S7).

Taxonomic annotation analyses further demonstrated MetaflowX's advantages in genome recovery. GTDB-Tk classified 1 394 bins at the species level, with only 89 bins remaining unclassified—fewer than the 98 unclassified bins from nf-mag (Fig. 3D). Across taxonomic ranks, MetaflowX annotated more taxa at the genus (455 versus 434 for nf-mag, 405 for MetaWRAP) and species (455 versus 434 and 405, respectively) levels (Fig. 3C). Dereplication of representative genomes using Galah also confirmed higher species-level diversity (471 species representatives) and comparable strain-level recovery (504 strains) (Fig. 3E), indicating that MetaflowX retains both quality and taxonomic diversity. MetaflowX's resource efficiency remained evident in this module: it completed the binning and downstream analyses in 3.9 h and consumed 320 GB of disk space—compared to 6.0 h and 1.8 TB for nf-mag, and 136.2 h and 714 GB for MetaWRAP (Fig. 3F). Integrative performance analysis across bin quality metrics, taxonomic resolution, and computational requirements revealed that MetaflowX achieved the most balanced performance profile among evaluated workflows (Fig. 3G), delivering HQ MAG recovery with minimal computational overhead. These results support that the complementary integration of the SBO and PBO binning strategies provides a robust framework for future metagenomic binning analyses.

Assessment of metaflowx-refinereassembly (DR) efficacy in MAG quality

To enhance the quality of MAGs through targeted reassembly of binned genomes, MetaflowX incorporates a RefineRe-

assembly (DR) submodule as part of its Binning module. This process utilizes reads from either the sample with the highest abundance for the target bin (single assembly mode) or multiple closely related samples (co-assembly mode). The module selects target bins based on read count and depth profiles, extracting aligned reads from both the target bin genome and its closest GTDB [55–57] reference genome for subsequent reassembly. After removing shorter contigs and coverage outliers, the reassembled contigs undergo quality assessment using CheckM2. High quality bins (completeness > 90% and contamination < 5%) [38] are retained directly. Lower-quality bins are subjected to rebinning with SemiBin2 (single-coverage) and MetaBAT2 (multi-coverage), followed by consolidation using DAS Tool [58]. Bins failing to meet HQ standards undergo further refinement using Deepurify.

To assess the effectiveness of the DR workflow, we analyzed 23 metagenome samples from a human early-life cohort study [59]. The initial analysis generated 364 genomic bins, from which 111 bins (with 50–90% completion and < 10% contamination) were selected for reassembly optimization. Following reassembly, 18 of these bins were elevated to HQ status. The remaining 93 bins underwent further refinement: 39 reached MQ, while 54 were downgraded due to excessive contamination. Subsequent re-binning of these 93 bins identified an additional 23 HQ bins. However, 13 bins exhibited reduced completeness and were excluded from the refinement process. Contaminant sequence removal using Deepurify on the remaining 49 bins resulted in 17 being upgraded to HQ, with the remaining 32 also showing improved quality compared to their initial versions. Notably, 88.3% of the bins demonstrated quality improvements, with 52.5% being upgraded to HQ, underscoring the effectiveness of the DR pipeline (Fig. 4A).

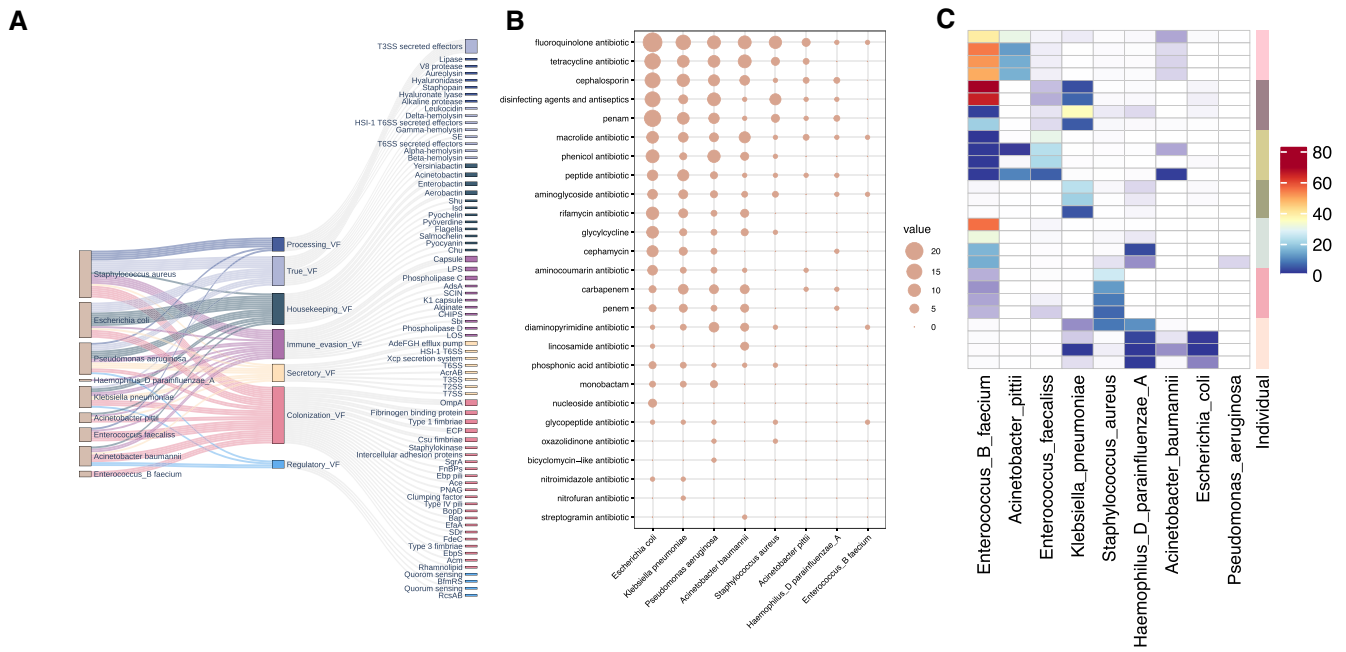


Figure 5. Bacterial virulence factors and antimicrobial resistance in intestinal obstruction pathogens. **(A)** Sankey diagram depicting virulence factor genes (VFGs) in bacterial species. Columns represent species carrying VFGs, VFG categories, and VFG types. **(B)** Distribution of antimicrobial resistance genes (ARGs) across bacterial species, organized by drug class. Circle size indicates the number of ARGs per drug class within each species. **(C)** Heatmap illustrating the relative abundance of species harboring VFGs and ARGs. Right-side annotations highlight individual patient variations.

The DR module significantly improved bin quality in single assembly mode, elevating 38.73% of MQ bins to high quality and generating 7 near-complete bins (>95% completeness, <1% contamination). The reassembly process increased overall bin set completeness by a median of 5.64% and reduced contamination by a median of 53.17% (Fig. 4C and D). In co-assembly mode (Fig. 4B), the number of HQ bins increased by 27%, with a 3.63% improvement in completeness and a 27.8% reduction in contamination (Fig. 4C and D, Supplementary Table S6). However, DR failed to enhance initially unclassified MAGs, as only 1 out of 7 bins lacking species-level GTDB annotation showed quality improvement. These results indicated that using the co-assembly method is more prone to introducing contaminants, which can negatively impact bin quality. Conversely, single assembly mode results in lower contamination rates while maintaining comparable completeness levels (Fig. 4E and F, Supplementary Fig. S8, Supplementary Table S7). MetaflowX therefore recommends single assembly mode as the default. Collectively, these results demonstrated that MetaflowX effectively improves both the completeness and purity of bins in both co-assembly and single-sample assembly modes.

Simultaneous identification of pathogenic and antibiotic-resistant bacteria

The identification of potentially pathogenic antibiotic-resistant bacteria (PARB) is critical for microbiological risk assessment. Conventional tools, such as MetaVF [60] or RGI [50], either estimate gene abundance based on reads or identify specific genes from contigs, but do not perform both tasks simultaneously. In contrast, MetaflowX can concurrently construct gene catalogs and genome sets (via the Geneset [4] and Binning [5] module) while maintaining detailed records of gene-contig-bin relationships. This capability enables the

rapid and simultaneous identification of virulence factors (VFGs) and antibiotic resistance genes (ARGs) in PARBs from metagenomic data. Additionally, it allows for the efficient calculation of their abundances at both the gene and bin levels.

To assess MetaflowX's capacity for detecting PARBs, we analyzed 28 metagenomic samples collected from four distinct intestinal fluid sites across seven patients with varying degrees of intestinal obstruction. Our analysis revealed the potential pathogens carrying substantial numbers of VFGs and ARGs (Fig. 5, Supplementary Tables S8, S9). *Escherichia* (*E.*) *coli* and *Klebsiella* (*K.*) *pneumoniae* emerged as the most prominent pathogens, with the following counts: VFGs (18 in *E. coli* and 11 in *K. pneumoniae*) and ARGs (51 in *E. coli* and 30 in *K. pneumoniae*). *K. pneumoniae* was detected in four patients, while *E. coli* was elevated in only one. Although *Enterococcus faecium* was dominant, it contained relatively few VFGs and ARGs (3 and 4, respectively). *Pseudomonas* (*P.*) *aeruginosa*, detected at low abundance in one patient, carried 16 VFGs and 23 ARGs. Prevalent VFGs were associated with colonization, housekeeping, and immune evasion, including siderophore systems (aerobactin, salmochelin, yersiniabactin, and enterobactin), structural components (OmpA and LPS), and secretion systems (Type III (T3SS) and Type VI (T6SS), along with their respective effectors). All potential pathogens carried fluoroquinolone resistance genes; *E. coli*, *K. pneumoniae*, and *P. aeruginosa* primarily exhibited resistance to fluoroquinolones, tetracyclines, and cephalosporins.

These findings underscore the significant presence of ARGs in potential pathogens, highlighting the need for judicious antibiotic use. MetaflowX offers HQ MAGs, enabling the simultaneous identification and quantification of both VFGs and ARGs while streamlining the traditional multi-step analysis process. This optimization reduces the time clinicians require to obtain critical PARB information, thereby facilitating more

efficient and accurate microbiological risk assessments in clinical practice.

Discussion

MetaflowX is designed to address the computational and structural challenges of large-scale metagenomic analyses. It imposes no restrictions on sample number or study design, supporting both small-scale exploratory analyses and production-level pipelines. By simplifying execution and integrating multi-step processes into a cohesive workflow, MetaflowX lowers the technical barrier for metagenomic analysis and reduces the likelihood of manual error. Standardized processing and comprehensive logging enable reproducible analyses across projects and facilitate methodological consistency in large-scale studies.

Our benchmarking results show that it achieves substantial gains in runtime and storage efficiency—completing full analyses up to 14 times faster and using 38% less disk space than comparable workflows. These improvements are enabled by Nextflow-based task orchestration, intermediate file cleanup, and support for chunked execution in high-complexity steps. Compared to other workflows (e.g. MetaWRAP, nf-mag, and anvi'o), MetaflowX offers higher compatibility with third-party results and greater flexibility in customizing execution paths. Its fine-grained modularity allows users to selectively run only essential components, avoiding unnecessary re-computation and enhancing reusability. In contrast, nf-mag's limited module granularity restricts fine control over execution, while MetaWRAP's reliance on fixed tool and database versions hinders integration of updated resources and methods. Anvi'o results are tightly coupled to its internal formats and tools, limiting interoperability with external results or pipelines. Importantly, MetaflowX supports large-scale analysis on moderate computing infrastructure (e.g. 32-core, 64GB RAM), minimizing reliance on high-performance hardware. In both benchmarking experiments and real-world projects, MetaflowX demonstrated high efficiency and low resource cost, highlighting its potential to promote more sustainable and reproducible metagenomic research practices [61].

In the context of MAG recovery, MetaflowX demonstrated the highest yield of HQ and MQ bins across multiple CAMI datasets. However, PBO's stringent consensus criteria may lead to reduced completeness for some genomes. Additionally, despite high recovery rates, not all bins could be taxonomically classified at the species level, indicating ongoing limitations in reference databases and strain-level diversity. While the DR module improved genome quality by increasing completeness and reducing contamination for over half of MQ bins, full recovery of all microbial genomes remains constrained by both biological complexity and computational limitations. These findings highlight the trade-offs between MAGs stringency, completeness, and annotation depth in metagenomic workflows.

The reduced costs of metagenomic sequencing have significantly advanced large-scale efforts to construct unique bacterial genetic repertoires, including a substantial proportion of genes with no known homologs—often termed 'microbial dark matter' [7, 62–65]. Also, recent discoveries of novel antibiotics [66] and antimicrobial peptides [67] from this 'dark matter' highlight its potential. MetaflowX's dual-track approach maximizes this opportunity by enabling comprehen-

sive gene cataloging alongside complete bacterial genome reconstruction. A key advantage lies in the efficient utilization of sequencing data, ensuring that even samples lacking complete genomes contribute valuable genetic information to the collective gene repository. Moreover, the generation of standardized metagenomic outputs further enables the effective application of Artificial Intelligence methods to mine and model microbial genetic diversity.

We demonstrated MetaflowX's capacity to rapidly identify both ARGs and VFGs across gene catalogs and assembled genomes, thereby facilitating the detection of potentially pathogenic, antibiotic-resistant bacteria. HQ genomes and genetic elements obtained through this reference-free approach provide valuable benchmarks for subsequent research, broadening our understanding of previously uncharacterized microbes. Furthermore, the vast diversity of uncultured bacterial species represents a rich resource for uncovering novel metabolic pathways and host-microbe interactions [68, 69]. Mining this 'microbial dark matter' could ultimately yield promising structural templates for next-generation antibiotics and antimicrobial peptides, helping to address the critical challenge of antimicrobial resistance [66, 70].

MetaflowX has certain limitations that should be considered. First, MetaflowX currently focuses on short-read sequences, which remains widely used due to its cost-effectiveness. Nonetheless, long-read (LR) sequencing technologies, such as PacBio and Nanopore, are increasingly critical for improving genome completeness and reducing contamination [71, 72]. Second, although integrating multiple binning algorithms improves the accuracy of bin recovery, it also introduces processing redundancy. For instance, tasks such as gene prediction with Prodigal and quality assessment with CheckM2 are repeated for each binning result. These tasks cannot be optimized through parameter adjustments to avoid redundancy, as they cannot be pre-executed or deferred post-execution. Third, in the DR submodule, applying the co-assembly approach to reassemble MAGs can introduce contamination and increase heterogeneity, particularly in highly diverse or structurally variable species [58]. Therefore, this module is primarily recommended for use with longitudinal data from the same individual [6]. Lastly, while MetaflowX supports microbial annotation using MetaPhlAn, Kraken, and GTDB, it does not provide comprehensive mapping relationships across these or other databases.

The future development of MetaflowX encompasses several directions. A key priority is to expand compatibility with long-read sequencing technologies by integrating specialized tools such as metaFlye [73], metaMDBG [74], HyLight [75], hifiasm-meta [76], MetaMaps [77], OPERA-MS [78], and Strainy [79], which will enable the generation of higher-quality MAGs [80, 81]. Incorporating advanced machine learning frameworks, particularly protein structure prediction tools like AlphaFold [82], offers the potential to deepen functional characterization of unannotated sequences within "microbial dark matter," thereby providing unprecedented insights into microbial diversity and function [83–87]. Additionally, future developments will focus on strain-level [88] identification and pangenome analysis [89–91], facilitating the exploration of microbial metabolic pathways and the rapid identification of prime candidates for genetic or biochemical studies aimed at elucidating their mechanisms of action [69, 92, 93]. We also aim to enhance MetaflowX's abil-

ity to recover MAGs from low-biomass environments, thereby broadening its applicability across diverse ecological contexts.

In conclusion, MetaflowX provides a flexible framework that enables users to customize and tailor their analyses. Its design accommodates a wide range of applications, including medical diagnostics, environmental monitoring, and biotechnology. The workflow is robustness and standardization, which enhance the reproducibility of end-to-end data processing in metagenomic research. This streamlined approach facilitates the sharing and reuse of metagenomic data across different studies, reducing the computational burden of reanalyzing raw data through various methods [94].

Acknowledgements

Not applicable.

Author contributions: Y.X. (Conceptualization [lead], Funding acquisition [lead], Methodology [equal], Supervision [equal], Writing – original draft [supporting], L.L. (Formal Analysis [lead], Methodology [equal], Software [supporting], Supervision [equal], Visualization [lead], Writing – original draft [lead], Writing – review & editing [equal], X.W. (Methodology [equal], Supervision [lead], Z.C. (Investigation [equal], Software [equal], J.L. (Software [supporting], Y.Y. (Software [lead], H.X. (Formal Analysis [supporting], Software [equal], Z.D. (Investigation [equal], X.H. (Visualization [supporting], S.L. (Investigation [supporting], Z.W. (Investigation [supporting], X.X. (Investigation [supporting], C.D. (Writing – review & editing [equal], Q.C. (Funding acquisition [supporting], Writing – review & editing [equal], Q.F. (Funding acquisition [supporting], Writing – original draft [supporting], Writing – review & editing [equal].

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

All authors declare no competing interests.

Funding

This research was supported by the National Key R&D Program of China (No. 2022YFA1304100), the National Natural Science Foundation of China (No. 82 270 980, No. 82071122, and No. 82 202 539), the National Science and Technology Major Program (2023ZD0501406), the National Young Scientist Support Foundation (2019), Excellent Young Scientist Foundation of Shandong Province (No. ZR2021JQ29), Taishan Young Scientist Project of Shandong Province (2019), Periodontitis innovation team of Jinan City (2021GXRC021), Major Innovation Projects in Shandong Province (No. 2021SFGC0502), Oral Microbiome Innovation Team of Shandong Province (No. 2020KJK001), Shandong Province Key Research and Development Program (No. 2021ZDSYS18), horizontal cooperation project with Shenzhen 01 Life Institute (#202412A001, #202112E401). Funding to pay the Open Access publication charges for this article was provided by the National Science and Technology Major Program (2023ZD0501406).

Data availability

No new sequencing data were generated for this study. The CAMI II benchmarking dataset is publicly available at <https://frrl.publisso.de/data/>. The real-world metagenomic sequencing data used for assessment are available at NCBI BioProject under the accessions PRJNA698986, PRJNA658385, and PRJNA398089. Detailed Sequence Read Archive (SRA) information is provided in the Supplementary Table S4. The MetaflowX workflow is openly accessible on GitHub (<https://github.com/01life/MetaflowX>) and on Zenodo (<https://doi.org/10.5281/zenodo.14166584>), and is freely available under the MIT license. The analysis code, intermediate results and source data can be found on Github (https://github.com/01life/MetaflowX_V1_DATA) and on Zenodo (<https://doi.org/10.5281/zenodo.14166606>).

References

1. Nayfach S, Roux S, Seshadri R *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499–509. <https://doi.org/10.1038/s41587-020-0718-6>
2. 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>
3. Coelho LP, Alves R, del Río ÁR *et al.* Towards the biogeography of prokaryotic genes. *Nature* 2022;601:252–6. <https://doi.org/10.1038/s41586-021-04233-4>
4. Almeida A, Mitchell AL, Boland M *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 2019;568:499–504. <https://doi.org/10.1038/s41586-019-0965-1>
5. Jin H, Quan K, He Q *et al.* A high-quality genome compendium of the human gut microbiome of Inner Mongolians. *Nat Microbiol* 2023;8:150–61. <https://doi.org/10.1038/s41564-022-01270-1>
6. Carter MM, Olm MR, Merrill BD *et al.* Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* 2023;186:3111–24. <https://doi.org/10.1016/j.cell.2023.05.046>
7. Carlino N, Blanco-Míguez A, Punčochář M *et al.* Unexplored microbial diversity from 2, 500 food metagenomes and links with the human microbiome. *Cell* 2024;187:5775–95. <https://doi.org/10.1016/j.cell.2024.07.039>
8. Mei Z, Wang F, Bhosle A *et al.* Strain-specific gut microbial signatures in type 2 diabetes identified in a cross-cohort analysis of 8, 117 metagenomes. *Nat Med* 2024;30:2265–76. <https://doi.org/10.1038/s41591-024-03067-7>
9. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 2004;38:525–52. <https://doi.org/10.1146/annurev.genet.38.072902.091216>
10. Meyer F, Lesker TR, Koslicki D *et al.* Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat Protoc* 2021;16:1785–801. <https://doi.org/10.1038/s41596-020-00480-3>
11. 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>
12. Saheb Kashaf S, Almeida A, Segre JA *et al.* Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat Protoc* 2021;16:2520–41. <https://doi.org/10.1038/s41596-021-00508-2>
13. Uritskiy GV, Diruggiero J, Taylor J. MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems. *Microbiome* 2018;6:158.

14. McIver LJ, Abu-Ali G, Franzosa EA *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* 2018;34:1235–7. <https://doi.org/10.1093/bioinformatics/btx754>
15. Beghini F, McIver LJ, Blanco-Míguez 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>
16. Kieser S, Brown J, Zdobnov EM *et al.* ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinf* 2020;21:257. <https://doi.org/10.1186/s12859-020-03585-4>
17. Krakau S, Straub D, Gourel H *et al.* nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genom Bioinform* 2022;4:lqac007. <http://doi.org/10.1093/nargab/lqac007>
18. Churchward B, Millet M, Bihouée A *et al.* MAGNETO: an automated workflow for genome-resolved metagenomics. *Msystems* 2022;7:e0043222. <https://doi.org/10.1128/msystems.00432-22>
19. Stamouli S, Beber ME, Normark T *et al.* nf-core/taxprofiler: highly parallelised and flexible pipeline for metagenomic taxonomic classification and profiling. *bioRxiv*, <https://doi.org/10.1101/2023.10.20.563221>, 23 October 2023, preprint: not peer reviewed.
20. Blanco-Míguez 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>
21. van Damme R, Hölzer M, Viehweger A *et al.* Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLoS Comput Biol* 2021;17:e1008716. <https://doi.org/10.1371/journal.pcbi.1008716>
22. Köster J, Mölder F, Jablonski KP *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33.
23. DI Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9. <https://doi.org/10.1038/nbt.3820>
24. 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;2019:e7359. <https://doi.org/10.7717/peerj.7359>
25. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;32:605–7. <https://doi.org/10.1093/bioinformatics/btv638>
26. 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>
27. Liu CC, Dong SS, Chen JB *et al.* MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 2022;10:46. <https://doi.org/10.1186/s40168-022-01237-8>
28. Pan S, Zhu C, Zhao XM *et al.* A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat Commun* 2022;13:2326. <https://doi.org/10.1038/s41467-022-29843-y>
29. Pan S, Zhao XM, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 2023;39, i21–9.
30. Nissen JN, Johansen J, Allesøe RL *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;39:555–60. <https://doi.org/10.1038/s41587-020-00777-4>
31. Wang Z, You R, Han H *et al.* Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat Commun* 2024;15:585.
32. Xia Y, Li X, Wu Z *et al.* Strategies and tools in illumina and nanopore-integrated metagenomic analysis of microbiome data. *Imeta* 2023;2:e72. <https://doi.org/10.1002/imt.2.72>
33. 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>
34. Rühlemann MC, Wacker EM, Ellinghaus D *et al.* MAGScoT: a fast, lightweight and accurate bin-refinement tool. *Bioinformatics* 2022;38:5430–3. <https://doi.org/10.1093/bioinformatics/btac694>
35. Zou B, Wang J, Ding Y *et al.* A multi-modal deep language model for contaminant removal from metagenome-assembled genomes. *Nat Mach Intell* 2024;6:1245–55. <https://doi.org/10.1038/s42256-024-00908-5>
36. Du Y, Sun F. MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. *Nat Commun* 2023;14:1245–55. <https://doi.org/10.1038/s41467-023-41209-6>
37. Qiu Z, Yuan L, Lian CA *et al.* BASALT refines binning from metagenomic data and increases resolution of genome-resolved metagenomic analysis. *Nat Commun* 2024;15:2179.
38. Bowers RM, Kyrpides NC, Stepanauskas R *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>
39. Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–90. <https://doi.org/10.1093/bioinformatics/bty560>
40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20. <https://doi.org/10.1093/bioinformatics/btu170>
41. 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>
42. 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>
43. 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>
44. 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>
45. 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>
46. Tatusov RL, Koonin EV, Lipman DJ. A Genomic Perspective on Protein Families. *Science* 1997;278:631–7. <https://doi.org/10.1126/science.278.5338.631>
47. Galperin MY, Wolf YI, Makarova KS *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;49:D274–81. <https://doi.org/10.1093/nar/gkaa1018>
48. Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* 2013;1:13–4.
49. Drula E, Garron ML, 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>
50. 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>

51. Liu B, Zheng D, Zhou S *et al.* VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* 2022;50:D912–7. <https://doi.org/10.1093/nar/gkab1107>
52. Han H, Wang Z, Zhu S. Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes. *Nat Commun* 2025;16:2865.
53. Eren AM, Esen OC, Quince C *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015;2015:e1319. <https://doi.org/10.7717/peerj.1319>
54. Eren AM, Kiehl E, Shaiber A *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2020;6:3–6. <https://doi.org/10.1038/s41564-020-00834-3>
55. Parks DH, Chuvochina M, Rinke C *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–94. <https://doi.org/10.1093/nar/gkab776>
56. Chaumeil PA, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315–6. <https://doi.org/10.1093/bioinformatics/btac672>
57. Parks DH, Chuvochina M, Chaumeil PA *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 2020;38:1079–86. <https://doi.org/10.1038/s41587-020-0501-8>
58. Mattock J, Watson M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat Methods* 2023;20:1170–3. <https://doi.org/10.1038/s41592-023-01934-8>
59. 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>
60. Dong W, Fan X, Guo Y *et al.* An expanded database and analytical toolkit for identifying bacterial virulence factors and their associations with chronic diseases. *Nat Commun* 2024;15:8084. <https://doi.org/10.1038/s41467-024-51864-y>
61. Arif SJ, Graham SP, Abdill RJ *et al.* Analyzing human gut microbiome data from global populations: challenges and resources. *Trends Microbiol* 2025;25. <https://doi.org/10.1016/j.tim.2025.05.008>
62. Rinke C, Schwientek P, Sczyrba A *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;499:431–7. <https://doi.org/10.1038/nature12352>
63. Chen J, Jia Y, Sun Y *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* 2024;633:371–9. <https://doi.org/10.1038/s41586-024-07891-2>
64. Rodríguez del Río Á, Giner-Lamia J, Cantalapiedra CP *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* 2024;626:377–84. <https://doi.org/10.1038/s41586-023-06955-z>
65. Coelho LP, Alves R, del Río ÁR *et al.* Towards the biogeography of prokaryotic genes. *Nature* 2022;601:252–6. <https://doi.org/10.1038/s41586-021-04233-4>
66. Ma Y, Guo Z, Xia B *et al.* Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol* 2022;40:921–31. <https://doi.org/10.1038/s41587-022-01226-0>
67. Shukla R, Peoples AJ, Ludwig KC *et al.* An antibiotic from an uncultured bacterium binds to an immutable target. *Cell* 2023;186:4059–73. <https://doi.org/10.1016/j.cell.2023.07.038>
68. Wu G, Xu T, Zhao N *et al.* A core microbiome signature as an indicator of health. *Cell* 2024;187:6550–65. <https://doi.org/10.1016/j.tim.2025.05.008>
69. VanEvery H, Franzosa EA, Nguyen LH *et al.* Microbiome epidemiology and association studies in human health. *Nat Rev Genet* 2023;24:109–24. <https://doi.org/10.1038/s41576-022-00529-x>
70. Kingwell K. Microbial 'dark matter' yields new antibiotic. *Nat Rev Drug Discov* 2023;22:872. <https://doi.org/10.1038/d41573-023-00156-z>
71. Agostinho DP, Fu Y, Menon VK *et al.* Unveiling microbial diversity: harnessing long-read sequencing technology. *Nat Methods* 2024;21:954–66. <https://doi.org/10.1038/s41592-024-02262-1>
72. Saheb Kashaf S, Almeida A, Segre JA *et al.* Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat Protoc* 2021;16:2520–41. <https://doi.org/10.1038/s41596-021-00508-2>
73. 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>
74. 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>
75. Kang X, Zhang W, Li Y *et al.* HyLight: strain aware assembly of low coverage metagenomes. *Nat Commun* 2024;15:8665. <https://doi.org/10.1038/s41467-024-52907-0>
76. Feng X, Cheng H, Portik D *et al.* Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat Methods* 2022;19:671–4. <https://doi.org/10.1038/s41592-022-01478-3>
77. Diltthey 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>
78. 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>
79. 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>
80. Kim CY, Ma J, Lee I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat Commun* 2022;13:2034–43. <https://doi.org/10.1038/s41467-022-34149-0>
81. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* 2020;38:701–7. <https://doi.org/10.1038/s41587-020-0422-6>
82. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;630:493–500. <https://doi.org/10.1038/s41586-024-07487-w>
83. Zheng W, Wuyun Q, Li Y *et al.* Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* 2024;21:279–89. <https://doi.org/10.1038/s41592-023-02130-4>
84. Baltoumas FA, Karatzas E, Liu S *et al.* NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res* 2024;52:D502–12. <https://doi.org/10.1093/nar/gkad800>
85. Hoarfrost A, Aptekmann A, Farfánuk G *et al.* Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat Commun* 2022;13:2606. <https://doi.org/10.1038/s41467-022-30070-8>
86. Miller D, Stern A, Burstein D. Deciphering microbial gene function using natural language processing. *Nat Commun* 2022;13:5731. <https://doi.org/10.1038/s41467-022-33397-4>
87. Santos-Júnior CD, Torres MDT, Duan Y *et al.* Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell* 2024;187:3761–78. <https://doi.org/10.1016/j.cell.2024.05.013>
88. Truong DT, Tett A, Pasolli E *et al.* Microbial strain-level population structure and genetic diversity from metagenomes.

- Genome Res* 2017;27:626–38.
<https://doi.org/10.1101/gr.216242.116>
89. Beavan A, Domingo-Sananes MR, McInerney JO. Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. *Proc Natl Acad Sci USA* 2024;121:e2304934120.
<https://doi.org/10.1073/pnas.2304934120>
 90. Shoer S, Reicher L, Zhao C *et al.* Pangenomes of human gut microbiota uncover links between genetic diversity and stress response. *Cell Host Microbe* 2024;32:1744–57.
<https://doi.org/10.1016/j.chom.2024.08.017>
 91. Heumos S, Heuer ML, Hanssen F *et al.* Cluster-efficient pangenome graph construction with nf-core/pangenome. *Bioinformatics* 2024;40:btac609.
<https://doi.org/10.1093/BIOINFORMATICS/BTAE609>
 92. 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>
 93. Liu S, Moon CD, Zheng N *et al.* Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* 2022;10:76.
<https://doi.org/10.1186/s40168-022-01272-5>
 94. Huttenhower C, Finn RD, McHardy AC. Challenges and opportunities in sharing microbiome data and analyses. *Nat Microbiol* 2023;8:1960–70.
<https://doi.org/10.1038/s41564-023-01484-x>