Nucleic Acids Research, 2025, gkaf1234
https://doi.org/10.1093/nar/gkaf1234
Database issue

MicrobialScope: an integrated genomic resource with rich annotations across bacteria, archaea, fungi, and viruses

Xikang Feng[1,2,†], **Yinhu Li**[3,4,5,†], **Jieyi Zheng**[1,†], **Xuhua Chen**[3,4,5,†], **Shuo Yang**[6], **Yu Chen**[3,4,5,*], **Shuai Cheng Li**[6,7,*]

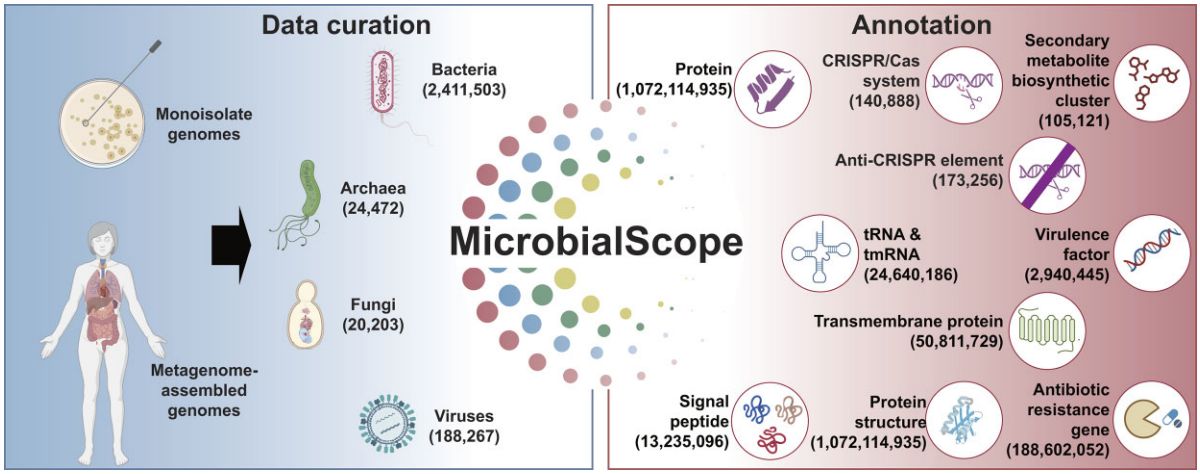
[1]School of Software, Northwestern Polytechnical University, Xi'an 710072, China
[2]Research & Development Institute, Northwestern Polytechnical University, Shenzhen 518063, China
[3]Shenzhen–Hong Kong Institute of Brain Science, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[4]The Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[5]SIAT–HKUST Joint Laboratory for Brain Science, Chinese Academy of Sciences, Shenzhen 518055, China
[6]Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China
[7]City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China

*To whom correspondence should be addressed: Email: shuaicli@cityu.edu.hk
Correspondence may also be addressed to Yu Chen. Email:yu.chen@siat.ac.cn
†These authors contributed equally to this work.

Abstract

Microorganisms, including bacteria, archaea, fungi, and viruses, are the most taxonomically diverse and ecologically dominant life forms on Earth, playing critical roles in ecosystems, human health, and industrial applications. While existing microbial databases such as BV-BRC and IMG archive both monoisolate and metagenome-assembled genomes (MAGs) across domains, challenges remain in standardized, multi-level annotations and interactive tools for all microbial groups. Here, we present MicrobialScope (https://microbial.deepomics.org/), a comprehensive microbial genomic platform that integrates large-scale genome collections, multilevel annotations, and interactive visualizations. MicrobialScope harbors 2 411 503 bacterial, 24 472 archaeal, 20 203 fungal, and 188 267 viral genomes derived from both monoisolate assemblies and MAGs. Integrating 15 state-of-the-art bioinformatics tools and 10 specialized databases, MicrobialScope provides extensive annotations encompassing basic genomic features, genomic element prediction (e.g., genes, tRNAs, tmRNAs, CRISPR–Cas and anti-CRISPR elements, secondary metabolite biosynthetic clusters, signal peptides, and transmembrane proteins), and functional and structural annotations. This includes 1 072 114 935 proteins with diverse annotations, 24 640 186 tRNAs and tmRNAs, 140 888 CRISPR–Cas systems, 173 256 anti-CRISPR elements, 105 121 secondary metabolite biosynthetic clusters, 13 235 096 signal peptides, and 50 811 729 transmembrane proteins. In addition, MicrobialScope offers unrestricted access to all data resources, interactive visualization tools, and built-in online analytical modules for intuitive exploration and comparative analysis. With its extensive genome collection, comprehensive annotations, and user-friendly interface, MicrobialScope serves as a scalable platform to advance genome research across diverse microbial domains.

Graphical abstract



Received: August 14, 2025. **Revised:** October 16, 2025. **Accepted:** October 16, 2025

Introduction

Microorganisms, including bacteria, archaea, fungi, and viruses, constitute the most taxonomically diverse and ecologically dominant group of organisms on Earth [1–3]. Microbial communities exhibit extraordinary biodiversity and are established across nearly all ecological niches [4–6]. In addition to their immense biodiversity, microorganisms harbor a vast repertoire of functional genetic elements [7], such as virulence factors [8], antibiotic resistance genes [9], and secondary metabolite biosynthetic clusters [10]. These genetic elements not only support microbial adaptation to extreme and competitive environments [6] but also play critical roles in host physiology as well as human health and disease [11].

Microbial genomes harbor rich functional elements, which facilitates investigation into their ecological adaptability, evolutionary trajectories, and impacts on host health [6, 11–13]. Although the distribution of genetic components varies across bacteria, archaea, fungi, and viruses, several classes of functional elements are of particular importance. Protein-coding genes and transfer RNAs (tRNAs) are fundamental for maintaining essential cellular processes and supporting protein translation across all microbial taxa, while transfer-messenger RNAs (tmRNAs) perform similar roles primarily in prokaryotes [14]. CRISPR–Cas systems and anti-CRISPR elements contribute to microbial immune defense and counter-defense mechanisms [15], shaping host–virus dynamics and promoting microbial evolution. Signal peptides and transmembrane proteins facilitate protein secretion, membrane transport, and environmental sensing [16], which are critical for microbial survival in diverse complex niches. Virulence factors mediate host colonization and pathogenesis [8], while antibiotic resistance genes are widely distributed among bacteria and some fungi [9], contributing to microbial pathogenesis and clinical challenges. In addition, secondary metabolite biosynthetic clusters produce bioactive compounds [10], which promote interspecies interactions and chemical communication. The application of large-scale language models [17] has significantly improved protein structure prediction, thereby enhancing the analysis of microbial functionality at the molecular level. These functional elements collectively represent a rich genetic resource crucial for elucidating microbial evolution and leveraging their roles in host health [7, 11, 18].

Several microbial genome repositories have advanced microbial research, yet gaps remain in achieving comprehensive, standardized, and interactive resources. Notable databases, such as NCBI RefSeq [19], EBI MGnify [20], BV-BRC [21], and IMG [22], curate large-scale genomic data, including monoisolate genomes and metagenome-assembled genomes (MAGs), with standardized annotations and taxonomic frameworks. For example, IMG integrates bacteria, archaea, fungi, and viruses with extensive MAGs, offering unified annotation pipelines and interactive visualizations, while BV-BRC provides standardized annotations and tools for bacteria and viruses. However, these resources vary in their coverage of all microbial domains, depth of functional annotations, and interactivity for comparative genomics.

To overcome these limitations, we developed MicrobialScope (https://microbial.deepomics.org/): an integrated genomic resource that offers extensive genome coverage, standardized genome annotation, and interactive data visualization. MicrobialScope incorporates high-quality genomes from NCBI RefSeq and GenBank as well as MAGs derived from diverse environments, such as the human gut, oral cavity, respiratory tract, skin, reproductive system, and other ecological niches. Each microbial genome in MicrobialScope contains comprehensive and standardized annotations that encompass basic features (e.g., genome length, assembly level, and taxonomic classification), genomic elements (e.g., genes, tRNAs, tmRNAs, CRISPR–Cas systems, and anti-CRISPR elements), functional annotations (e.g., virulence factors, antibiotic resistance genes, and secondary metabolite biosynthetic clusters), and protein structures predicted by a large protein language model. To facilitate data exploration, MicrobialScope also provides web-based tools for customizable querying, interactive genome browsing, and full data downloads. By integrating rich annotations with interactive access, MicrobialScope is a comprehensive and user-friendly platform for investigating microbial diversity, evolutionary biology, and host-associated functionality.

Materials and methods

Microbial collection and integration

To compile a comprehensive collection of microbial sequences, we gathered a wide array of single genomes and MAGs, integrating microbial genomes from four major domains: bacteria, archaea, fungi, and viruses. We collected single genomes from the NCBI RefSeq and GenBank databases (till 12 December 2024) [19, 23], along with MAGs deposited in these repositories from diverse environments, as well as MAGs reconstructed in our laboratory (e.g. RMGC). For the collected microbial genomes, assemblies underwent quality control based on genome size and GC content. Specifically, bacterial and archaeal genomes <0.1 Mb or >20 Mb, or with GC contents outside the 25%–75% range, were excluded. Fungal genomes <1 Mb or with GC contents < 25% or > 75% were removed. Viral genomes with sizes < 1 kb or > 2 Mb, or with GC contents outside 25%–75%, were filtered out. In addition, bacterial and archaeal genomes were further assessed using CheckM (v1.2.3) [24], and those with estimated completeness below 50% or contamination above 10% were excluded. After filtering, only high-quality single genomes and MAGs were retained, ensuring a consistent and reliable dataset for downstream analyses. We also identified duplicate sequences present across the four domains using MMseqs2 (v15.6f452) with the following clustering parameters to detect microbes with identical sequences: "–cov-mode 0 -c 1.0 -min-seq-id 1.0" [25]. The microbial IDs corresponding to these duplicate sequences were retained and are displayed adjacent to the "Genome List" page of the MicrobialScope website. Within these datasets, our collection covers not only genome sequences and their basic information (e.g., length, N50, assembly level (refers to the genome completeness as defined by NCBI, such as complete genome, chromosome, scaffold, or contig), etc.), but also their taxonomic information.

Annotation of microbial genes and genomic elements

For microorganisms without pre-existing genomic annotations in their source databases, we performed gene and genomic element annotations. Gene annotation differs substan-

tially between eukaryotes and prokaryotes because of their distinct gene structures. For bacterial, archaeal, and viral genomes, we used Prokka (v1.11) [26] for annotation, which integrates Prodigal (v1.6.3) [27] for identifying open reading frames and ARAGORN (v1.2.41) [28] for detecting tRNA and tmRNA genes. We annotated fungal genomes using the fungus-specific pipeline, Funannotate (v1.8.17) [29]. Funannotate is based on Evidence Modeler, which integrates multiple gene prediction inputs to produce consensus gene models. The supported ab initio gene predictors include Augustus (v3.5.0) [30], SNAP (2006-07-28) [31], GlimmerHMM (v3.0.4) [32], CodingQuarry (v2.0) [33], and GeneMark-ES/ET (v4.71_lic) [34–36]. Finally, we predicted tRNA genes using tRNAscan-SE (v2.0.12) [37].

We subsequently analyzed the CRISPR–Cas systems in prokaryotic genomes. We identified CRISPR arrays and their associated Cas genes in each plasmid genome using CRISPRCasTyper (v1.8.0) [38] and classified system subtypes based on a comprehensive analysis of Cas genes and CRISPR repeat sequences. We predicted anti-CRISPR proteins (Acrs) using AcrFinder (v2.0) [39], which combines sequence homology and guilt-by-association approaches for detection. We predicted signal peptides and their cleavage sites in bacterial, fungal, and archaeal proteins using SignalP 6.0 [40]. To identify transmembrane domains in membrane proteins across the four domains, we used TMHMM 2.0 [41], which applies Hidden Markov Models to capture structural complexity. We executed all tools with domain-specific parameter adjustments (Supplementary Table S1).

Functional annotation

For functional annotation of coding sequences, we used eggNOG-mapper (v2.1.12) [42] with the default parameters to perform rapid orthology assignments based on precomputed eggNOG (v5.0.2) clusters and phylogenies [43]. The resulting annotations included comprehensive matching and scoring information along with functional insights from multiple databases, such as Gene Ontology (GO) [44], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [45], the BiGG Database [46], Clusters of Orthologous Groups (COG) [47], and the Carbohydrate-Active EnZymes database (CAZy) [48] (Supplementary Table S2).

In addition, we conducted homology-based searches for microbial proteins using Diamond (v2.1.8.162) [49], referencing the Virulence Factor Database (VFDB) [50] for bacteria, archaea, and viruses, and the Database of Fungal Virulence Factors in Fungal Pathogens (DFVF) [51] for fungi. We identified virulence factors as matches exceeding 60% sequence identity and a 80% coverage threshold. We performed antibiotic resistance gene annotation for bacterial, archaeal, and fungal genomes using homology and single-nucleotide polymorphism modeling, referencing the Comprehensive Antibiotic Resistance Database (CARD) [52] with the parameters, "–include_loose" and "–include_nudge." For viruses, we identified their antibiotic resistance genes using AMRFinder-Plus (v4.0.23) [53] with the parameters, "–plus." Finally, we identified and annotated secondary metabolite biosynthetic gene clusters in bacterial, archaeal, and fungal genomes using antiSMASH (v7.1.0) [54], incorporating both domain-specific settings and the following additional parameters: "–asf –cc-mibig –cb-general –cb-knownclusters –cb-subclusters -pfam2go."

Protein structure prediction

We employed artificial intelligence-driven modeling approaches to predict protein structures and generate high-resolution 3D models. Specifically, we integrated ESMFold [17] into the web-based framework of MicrobialScope. To assess the confidence of the predicted structures, we utilized the predicted Local Distance Difference Test (pLDDT), which provides residue-level confidence scores. On each "Protein Detail" page, users can interactively visualize the 3D model, examine residue-specific pLDDT scores by hovering over individual residues, and download the corresponding CIF file containing structural coordinates and confidence metrics for further analysis.

Sequence alignment

To enable comparative analysis of protein coding sequences between user-submitted microbial genomes and those available in MicrobialScope, we employed BLASTP [55] to perform pairwise alignments of predicted proteins. The "Alignment Results" page provides an integrated view of each genome's gene predictions, associated functional annotations, and pairwise protein-level similarities, offering insights into their genetic relatedness. Moreover, all BLAST and alignment outputs are available for download, enabling users to perform downstream analysis.

Comparative analysis

To infer genetic relationships among microbial genomes, we implemented a two-stage comparative analysis pipeline. First, we calculated pairwise genome distances based on an alignment-free comparison that quantifies dissimilarity via Euclidean distance calculated from 6-mer frequency profiles using Alfpy [56]. Second, we applied the neighbor-joining algorithm to these distance matrices to construct a dendrogram representing genome-level similarity in a phylogeny-like structure. The resulting page provides a comparative tree available for download in PHY format.

Statistical analysis

To demonstrate the use of MicrobialScope, we retrieved coronavirus genomes for the case study, utilizing the filtering interface in the "Genome" section under the "Database" menu of MicrobialScope. We used the following settings: "Viruses" under "Microbe," "Monoisolate" under "Assembly Type," and "Complete Genome" under "Assembly Level." By entering the keyword "coronavirus" into the top-right search box using "Species" next to the search bar, we obtained 103 complete coronavirus genomes. To explore the phylogenetic relationships of these coronaviruses, we downloaded the corresponding metadata, genome sequences, and annotation files by clicking the "Download" button and constructed a whole-genome-based phylogenetic tree using the "Comparative Analysis" module under the "Analysis" menu. We subsequently employed iTOL (v7.2.1) to visualize key attributes of these coronaviruses [57], including taxonomic classification, host origin, gene counts, and the number of annotated antibiotic resistance genes. To demonstrate functional genome exploration, we further examined the genome annotation of a human-derived SARS-like coronavirus (i.e. GCA_031162155.1) and visualized the predicted 3D structures of its viral proteins by ESMFold embedded in MicrobialScope.

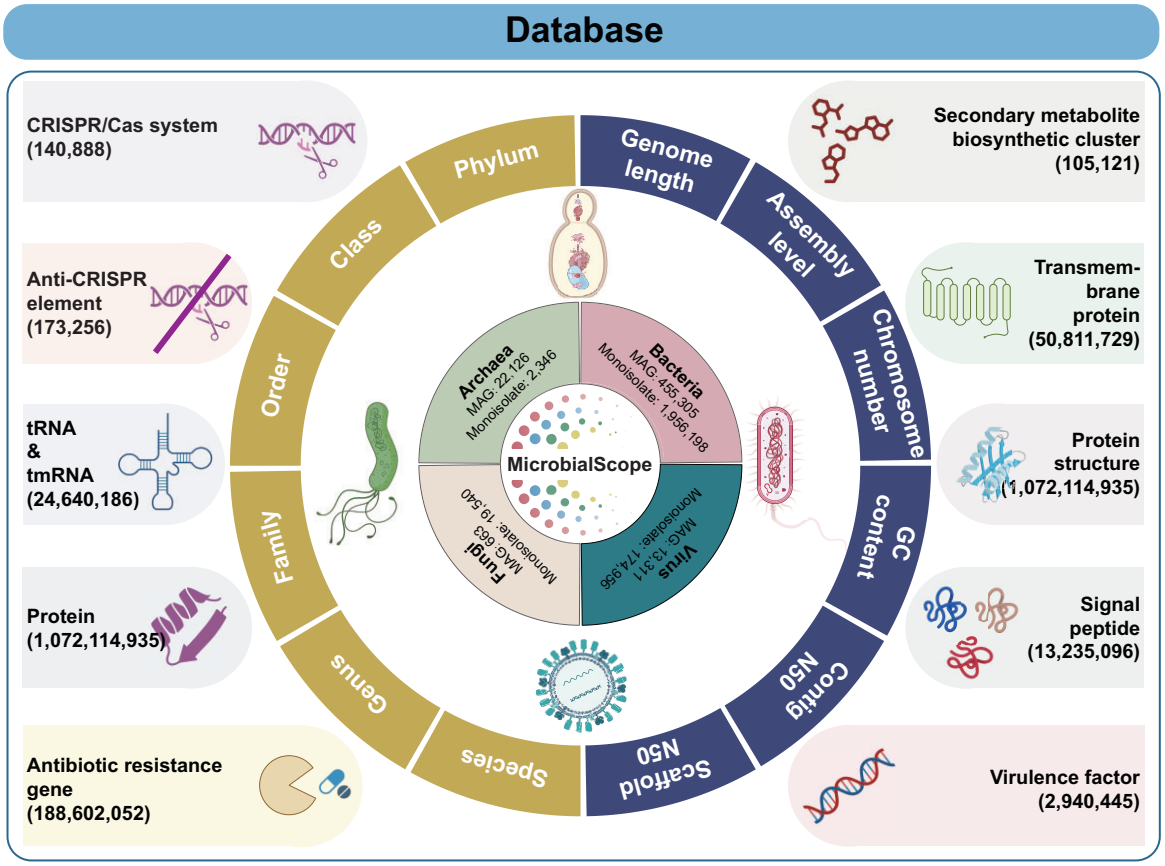


Figure 1. Overview of the microbial genomic resources in MicrobialScope. MicrobialScope integrates curated microbial genomes from bacteria, archaea, fungi, and viruses from public repositories, including both monoisolate genomes and MAGs. The platform offers standardized annotations covering six basic genomic features (i.e., genome length, GC content, assembly level, chromosome number, contig N50, and scaffold N50) and seven key genetic elements (i.e., genes, tRNAs, tmRNAs, CRISPR–Cas systems, anti-CRISPR elements, signal peptides, and transmembrane proteins). In addition, MicrobialScope provides diverse functional annotations and 3D protein structures for encoded genes.

Platform development

MicrobialScope is hosted on an Ubuntu 20.04.6 LTS server equipped with 1 TB memory and 90 TB storage. The platform's backend functionality is supported by an in-house framework consisting of Nginx, Django, PostgreSQL, and React+Next.js [58, 59]. We implemented all online data visualizations using Apache ECharts, D3.js, and Oviz [60]. We have also provided detailed tutorials on the platform to facilitate user navigation and utilization.

Results

Comprehensive and integrated microbial genomic resources in MicrobialScope

MicrobialScope integrates microbial genomes across four major domains—bacteria, archaea, fungi, and viruses—including monoisolate genomes and MAGs (Fig. 1). We retrieved monoisolate genomes from the NCBI RefSeq and GenBank repositories [19, 23], including 1 956 198 bacterial, 2346 archaeal, 19 540 fungal, and 174 956 viral genomes, accounting for 81.42% of the database. MicrobialScope incorporates MAGs derived from diverse environments, including the human gut [61, 62], oral cavity [63, 64], respiratory tract [65], skin [66], reproductive system [67], and other environments [68–70], contributing an additional 455 305 bacterial, 22 126 archaeal, 663 fungal, and 13 311 viral genomes, accounting for 18.58% of the database. To facilitate the usability and interoperability of MicrobialScope, we subjected all genomes to standardized multilevel annotation, including basic feature extraction, genomic element prediction, and functional and structural annotation.

MicrobialScope extracts six basic genomic features for each microbial genome: genome length, GC content, assembly level, chromosome number, contig N50, and scaffold N50 (Fig. 1). These features vary substantially across microbial domains. For example, bacterial genomes range from 100 225 to 18 931 163 bp with GC contents between 25% and 74.99%; archaeal genomes exhibit a comparable distribution (Supplementary Table S3). Fungal genomes tend to be larger, with lengths from 1 086 755 to 2 054 317 516 bp and GC contents between 25.08% and 68.47%. Meanwhile, viral genomes are generally smaller, ranging from 1000 to 1 908 524 bp with diverse GC contents. Based on assembly completeness, MicrobialScope classifies genomes into four tiers: complete genome, chromosome, scaffold, and contig. There are 66 134 bacterial, 759 archaeal, 1751 fungal, and 183 291 viral genomes with complete or chromosome-level assemblies. Among them, chromosome numbers range from 1 to 13 in bacteria, 1 to 9 in archaea, 1 to 43 in fungi, and 1 to 11 in viruses. As MAGs constitute a substantial proportion of MicrobialScope, they mainly remain at the scaffold or contig level, including 2 345 369 bacterial,

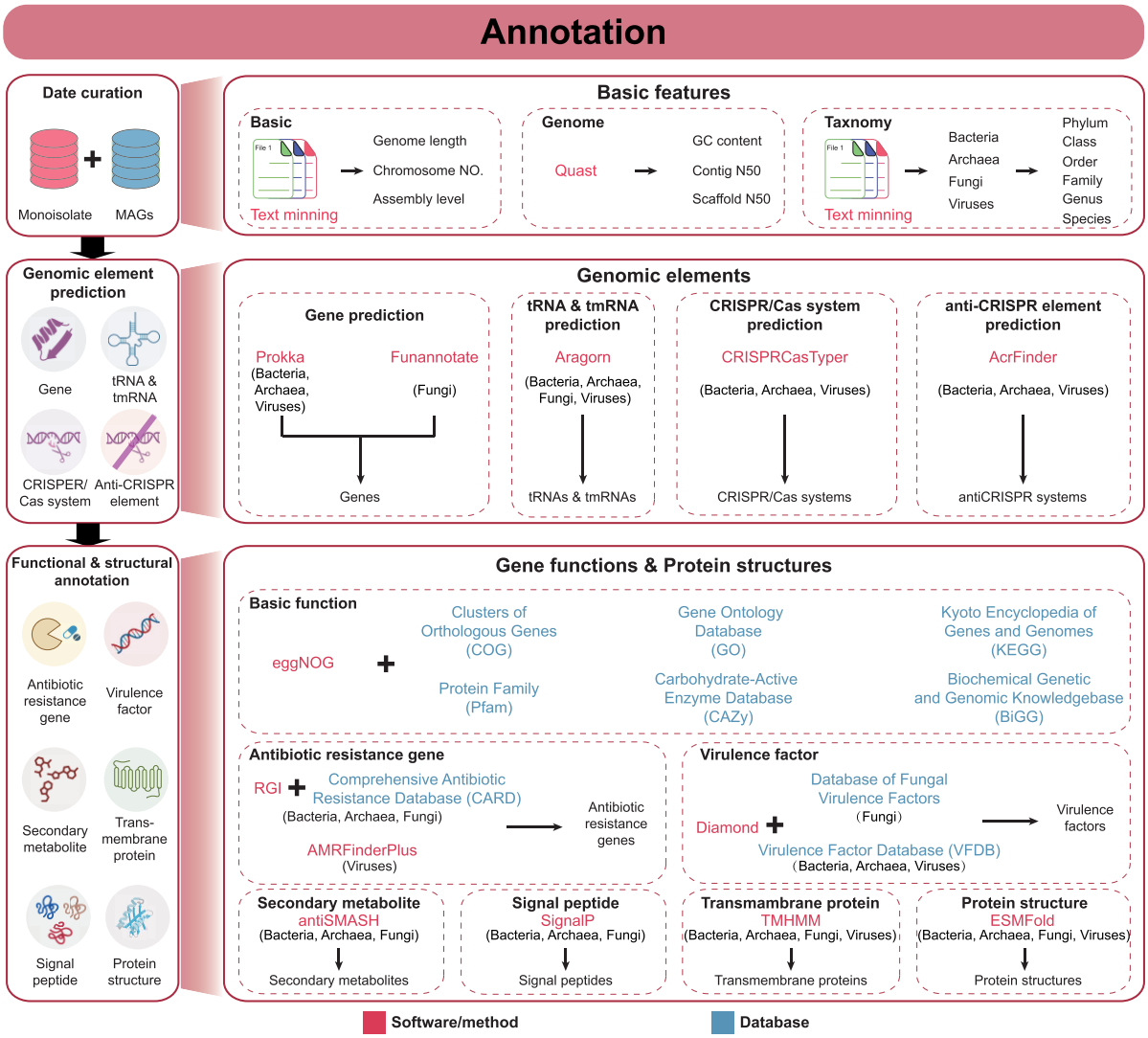


Figure 2. Overview of the annotation workflow in MicrobialScope. MicrobialScope employs 15 bioinformatics tools and 10 functional databases to perform standardized predictions of genetic elements and functional annotations for each microbial genome. The annotations are organized into 10 specialized datasets: Genome, Protein, tRNA and tmRNA, CRISPR–Cas System, Anti-CRISPR Element, Secondary Metabolite, Signal Peptide, Virulence Factor, Antibiotic Resistance Gene, and Transmembrane Protein.

23 713 archaeal, 18 452 fungal, and 4976 viral genomic sequences. The inclusion of genomic metadata enables users to assess genome quality, compare across taxa, and select appropriate genomes for downstream analysis.

MicrobialScope provides systematic annotations of core genomic elements that are tailored to the biological characteristics of each microbial domain (Fig. 2). For bacteria and archaea, MicrobialScope provides seven key features: genes, tRNAs, tmRNAs, CRISPR–Cas systems, anti-CRISPR elements, signal peptides, and transmembrane proteins. Using standardized bioinformatics pipelines, we identified 832 302 306 genes, 23 213 807 tRNAs and tmRNAs, 132 788 CRISPR systems, 2681 anti-CRISPR elements, 6 988 197 signal peptides, and 29 686 707 transmembrane proteins in bacterial genomes along with similar annotations for archaeal genomes (Supplementary Table S3). In fungi, MicrobialScope annotates four core features, resulting in the identification of 193 906 961 genes, 694 231 tRNAs and tmRNAs, 4 481 750 signal peptides, and 11 519 517 transmembrane proteins. For viruses, we annotated 6 593 428 genes, 88 568 tRNAs and tmRNAs, 77 CRISPR–Cas systems, 165 420 anti-CRISPR elements, and 987 537 transmembrane proteins (Supplementary Table S3). This extensive catalog of genomic components supports in-depth investigations of gene structure and function, facilitating the discovery of elements with potential clinical or biotechnological applications.

To enhance functional interpretation, MicrobialScope offers comprehensive functional annotations and protein structure prediction for each predicted gene (Fig. 2). We applied 10 well-established databases for functional annotations, including GO [44], KEGG [45], BiGG [46], COG [47], CAZy [48], VFDB [50], DFVF [51], CARD [52], Pfam [71], and MIBiG [72], which cover diverse aspects of microbial metabolism, pathogenicity, enzyme function, resistance mechanisms, and secondary metabolite biosynthesis. For each gene, MicrobialScope provides both database-specific annotations and category-level classifications (e.g. GO terms, COG functional groups, and KEGG pathways), enabling multidimensional exploration of gene functions. In addition, MicrobialScope employs the ESMFold deep learning frame-

work to predict protein structures at the atomic level [17]. Users can interactively visualize, manipulate, and download 3D protein models, facilitating structural studies and functional inference. This combination of functional and structural prediction offers users a rich resource for studying microbiology from both sequence- and structure-based perspectives.

Interactive and informative visualization in MicrobialScope

MicrobialScope features a well-structured, user-friendly web interface that facilitates querying, visualization, analysis, and data retrieval. The platform has eight main pages each designed to support navigation and data exploration: "Home," "Microorganism," "Database," "Analysis," "Workspace," "Download," "Tutorial," and "Contact us."

The "Home" page serves as the integrated entrance, featuring a search box that supports queries by genome ID or taxonomic name along with a dashboard-style summary of database contents. It displays the numbers of available genomes across the four microbial domains, the distribution of nine annotated features (e.g. genes, tRNAs–tmRNAs, and CRISPR–Cas systems), and two feature plots that offer quick insights into the overall annotation landscape. Clicking on any graph redirects users to the relevant dataset within the "Database" section.

The "Microorganism" page presents a comparative overview of monoisolate and MAG-derived genomes across the four microbial domains, including detailed counts of genomes and genetic elements. Users can further explore genome-level details by clicking the "Explore" button, which links to the corresponding "Genome" dataset within the "Database" section.

The "Database" section is the core of MicrobialScope, comprising 10 specialized annotation pages that correspond to distinct datasets: "Genome," "Protein," "tRNA and tmRNA," "CRISPR–Cas System," "Anti-CRISPR Element," "Secondary Metabolite Biosynthetic Cluster," "Signal Peptide," "Virulence Factor," "Antibiotic Resistance Gene," and "Transmembrane Protein" (Fig. 2). Each page includes a filtering panel on the left for category-based refinement and a search bar on the top right for targeted queries. All pages also contain "View Detail" and "Download" options, allowing users to access annotation metadata and download nucleotide or amino acid sequences as needed. Furthermore, a short tutorial pop-up window is automatically displayed for first-time visitors when clicking the "View Details" button under the "Genome" page.

To support downstream data exploration, the "Analysis" page provides six integrated online modules: "ORF Prediction and Protein Classification," "tRNA and tmRNA Prediction," "Virulence Factors and ARG Detection," "Transmembrane Protein Annotation," "Sequence Alignment," and "Comparative Analysis." All submitted tasks and results are automatically tracked in the "Workspace" section where users can monitor task status, access interactive results, and download outputs via the "View Result" function.

The "Download" page offers a list of processed datasets and annotation files. Additionally, the download API at "Download" page enables users to retrieve customized subsets of data, including Metadata, FASTA, GenBank, GFF, and Annotation Data, via flexible filtering options for efficient access to large-scale microbial genomic resources. While the "Tutorial" page provides detailed user guidance and step-by-step instructions to assist in usage. Finally, the "Contact us" page offers a direct channel for user feedback, facilitating continued platform development and community engagement.

Case study: genomic characterization of coronaviruses using MicrobialScope

To demonstrate the utility of MicrobialScope, we performed a case study focusing on the genomic and phylogenetic characteristics of coronaviruses using MicrobialScope. Using the 103 complete coronavirus genomes filtered from MicrobialScope, we found that the genome size ranged from 25 874 to 31 775 bp. These genomes contained 903 predicted genes (Fig. 3A). To explore the phylogenetic relationships of these coronaviruses, we downloaded the corresponding metadata, genome sequences, and annotation files from MicrobialScope. The results suggest that *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* form distinct clades, each associated with specific host ranges (Fig. 3A). Notably, both humans and bats primarily host *Alphacoronavirus* ($n = 8$ and 24, respectively) and *Betacoronavirus* ($n = 2$ and 29, respectively), while birds primarily host *Gammacoronavirus* ($n = 2$) and *Deltacoronavirus* ($n = 14$).

To illustrate functional genome exploration, we selected human-hosted SARS-like coronavirus WIV16 (i.e. GCA_031162155.1) for further analysis. The genome encodes 13 genes, including key components associated with viral replication and host infection (Fig. 3B). Using MicrobialScope's integrated protein structure prediction module, we obtained the structures of 6 key viral proteins, including replicase polyprotein 1a, replicase polyprotein 1ab, spike glycoprotein precursor, envelope small membrane protein, membrane protein, and nucleoprotein (Fig. 3C). These atomic-level protein structures provide valuable insights into domain architecture and offer potential targets for antiviral drug or vaccine design.

Thus, these results highlight MicrobialScope's capability of supporting end-to-end genomic investigation—from genome retrieval and annotation to evolutionary analysis and structural biology—demonstrating its potential as a valuable resource for studying microbial diversity, evolution, and pathogenesis.

Discussion

MicrobialScope is a comprehensive, integrative platform that offers large-scale, genomic resources across the four major microbial domains. It also features multiscale annotations, user interactivity, and analytical flexibility. MicrobialScope has four key features. First, MicrobialScope has an extensive data collection, incorporating over 2.6 million high-quality monoisolate genomes and MAGs, which are broadly representative of microbial diversity. Second, its comprehensive annotations support both genome- and structural-based research. Each genome undergoes standardized multiscale annotation, including basic genomic features, genomic element prediction, functional annotation, and protein structure prediction. Third, the platform provides informative, interactive visualization tools for exploring genomic features and annotation

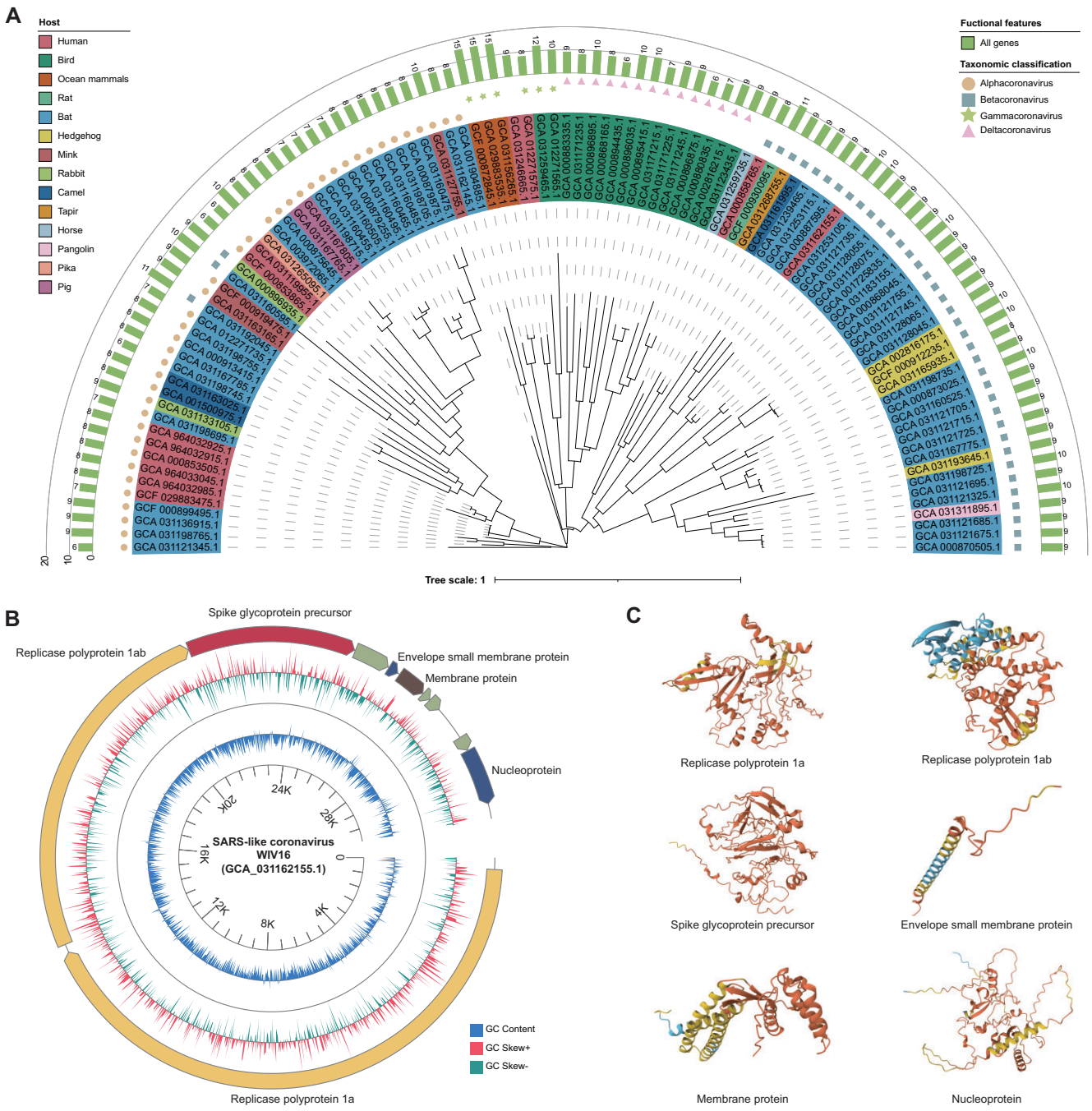


Figure 3. Case study of coronavirus genomic characterization using MicrobialScope. (**A**) Genomic and phylogenetic analysis of 103 complete coronavirus genomes retrieved from MicrobialScope. In the phylogenetic tree, node colors represent host origins of coronaviruses. Shapes indicate taxonomic classification: yellow circles for *Alphacoronavirus*, blue squares for *Betacoronavirus*, green stars for *Gammacoronavirus*, and pink triangles for *Deltacoronavirus*. Bar plots indicate numbers of predicted genes in each genome. (**B**) Genome structure of a representative SARS-like coronavirus (i.e. GCA_031162155.1). From outer to inner circles: genomic coordinates (kb), GC content, GC skew on the positive (red) and negative (green) strands, and annotated coding sequences. (**C**) Predicted 3D structures of six viral proteins from a representative SARS-like coronavirus (i.e. GCA_031162155.1) generated using the ESMFold model embedded in MicrobialScope.

results, enabling users to intuitively examine genomic structures and functional components. Fourth, MicrobialScope is open access with full data availability. MicrobialScope is freely accessible without registration and has full download support for all data resources.

MicrobialScope provides distinctive improvements over existing microbial genome databases and web tools in terms of data scale, annotation depth, and interactivity (Supplementary Table S4). First, MicrobialScope integrates over 2.6 million microbial genomes—including both monoisolates and MAGs—spanning bacteria, archaea, fungi, and viruses, whereas most existing resources emphasize either a specific domain (e.g. BV-BRC for pathogens [21], MBGD without viral genomes [73]) or focus primarily on monoisolates (e.g. RefSeq, IMG). Second, MicrobialScope delivers systematic and standardized annotations of func-

tional and structural genomic elements, covering underrepresented features often absent from other databases (e.g. NCBI RefSeq [19], EBI MGnify [20], PATRIC [74], IMG [22], MBGD [73], and CMR [75]), such as anti-CRISPR proteins, transmembrane proteins, and AI-driven protein structure predictions. Third, MicrobialScope provides user-friendly interface that supports real-time querying, interactive visualization, and data export, together with integrated analytical tools for ORF prediction, functional classification, sequence alignment, and comparative genomics. Collectively, these advances make MicrobialScope a powerful and scalable resource for microbial genome research and large-scale genetic element analysis.

MicrobialScope aims to support microbiological research by providing a broad, standardized, and sustainable genomic infrastructure. Accordingly, we will continue to update MicrobialScope in response to the demands of the research community. First, we will update MicrobialScope annually to incorporate newly released monoisolate genomes and MAGs and ensure its long-term utility. Second, we will implement additional bioinformatic and AI-assisted annotation modules to enhance online analysis, data mining, and biological interpretation. Third, we will strengthen community engagement through data submission, structured feedback, and collaborative development modules. Last, we will build secure, scalable infrastructure to support large-scale data processing, long-term storage, and efficient data retrieval.

In summary, MicrobialScope is a valuable microbial genomic resource that integrates comprehensive genomic data, rich multilevel annotations, and interactive analytical tools to empower systematic investigations of microbes and host–microbe interactions.

Acknowledgements

We thank the members of the Li Laboratory and Chen Laboratory for their helpful discussions and insights. We also thank the Shenzhen–Hong Kong Institute of Brain Science and the SIAT-HKUST Joint Laboratory of Brain Science for the instrument and technical support.

Author contribution: Xikang Feng (Software, Formal analysis, Methodology, Funding acquisition, Writing—original draft), Yinhu Li (Data curation, Formal analysis, Methodology, Visualization, Funding acquisition, Writing—original draft); Jieyi Zheng (Software, Formal analysis, Methodology, Writing—original draft), Xuhua Chen (Data curation, Formal analysis, Methodology, Visualization, Writing—original draft), Shuo Yang (Software, Writing—review & editing), Yu Chen (Conceptualization, Supervision, Funding acquisition, Writing—review & editing), and Shuai Cheng Li (Methodology, Conceptualization, Supervision, Funding acquisition, Writing—review & editing).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Young Collaborative Research Grant [C2004-23Y]; the Shenzhen Science and Technology Program [JCYJ20220818101201004]; the National Natural Science Foundation of China [32300527 and 32470695]; the Guangdong Basic and Applied Basic Research Foundation [2022A1515110784]; the Key-Area Research and Development Program of Guangdong Province [2023B0303040004]; the Basic Research Programs of Taicang, 2024 [TC2024JC43]; the Shenzhen–Hong Kong Institute of Brain Science; and the SIAT-HKUST Joint Laboratory of Brain Science. Funding to pay the Open Access publication charges for this article was provided by the Young Collaborative Research Grant [C2004-23Y].

Data availability

All data are freely available at https://microbial.deepomics.org/.

References

1. Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 1997;276:734–40. https://doi.org/10.1126/science.276.5313.734
2. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci* 2016;113:5970–5. https://doi.org/10.1073/pnas.1521291113
3. Averill C, Anthony MA, Baldrian P *et al.* Defending Earth's terrestrial microbiome. *Nat Microbiol* 2022;7:1717–25. https://doi.org/10.1038/s41564-022-01228-3
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. 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
6. Shu WS, Huang LN. Microbial diversity in extreme environments. *Nat Rev Microbiol* 2022;20:219–35. https://doi.org/10.1038/s41579-021-00648-y
7. Wilmes P, Simmons SL, Denef VJ *et al.* The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* 2008;33:109–32. https://doi.org/10.1111/j.1574-6976.2008.00144.x
8. Leitão JH. Microbial virulence factors. *International journal of molecular sciences* 2020.21:5320. https://doi.org/10.3390/ijms21155320
9. Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *science* 2009;325:1128–31. https://doi.org/10.1126/science.1176950
10. Dinglasan JLN, Otani H, Doering DT *et al.* Microbial secondary metabolites: advancements to accelerate discovery towards application. *Nat Rev Microbiol* 2025;23:338–54. https://doi.org/10.1038/s41579-024-01141-y
11. Ma Z, Zuo T, Frey N *et al.* A systematic framework for understanding the microbiome in human health and disease: from basic principles to clinical translation. *Signal Transd Targeted Ther* 2024;9:237. https://doi.org/10.1038/s41392-024-01946-6
12. Yang Y. Emerging patterns of microbial functional traits. *Trends Microbiol* 2021;29:874–82. https://doi.org/10.1016/j.tim.2021.04.004
13. Gevers D, Cohan FM, Lawrence JG *et al.* Re-evaluating prokaryotic species. *Nat Rev Microbiol* 2005;3:733–9. https://doi.org/10.1038/nrmicro1236

14. Jia X, He X, Huang C *et al.* Protein translation: biological processes and therapeutic strategies for human diseases. *Signal Transd Targeted Ther* 2024;9:44. https://doi.org/10.1038/s41392-024-01749-9
15. Kadkhoda H, Gholizadeh P, Kafil HS *et al.* Role of CRISPR–Cas systems and anti-CRISPR proteins in bacterial antibiotic resistance. *Heliyon* 2024;10:14. https://doi.org/10.1016/j.heliyon.2024.e34692
16. Hegde RS, Keenan RJ. The mechanisms of integral membrane protein biogenesis. *Nat Rev Mol Cell Biol* 2022;23:107–24. https://doi.org/10.1038/s41580-021-00413-2
17. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123–30. https://doi.org/10.1126/science.ade2574
18. Henry LP, Bruijning M, Forsberg SK *et al.* The microbiome extends host evolutionary potential. *Nat Commun* 2021;12:5141. https://doi.org/10.1038/s41467-021-25315-x
19. Goldfarb T, Kodali VK, Pujar S *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2025;53:D243–57. https://doi.org/10.1093/nar/gkae1038
20. 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
21. Olson RD, Assaf R, Brettin T *et al.* Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* 2023;51:D678–89. https://doi.org/10.1093/nar/gkac1003
22. Chen IMA, 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
23. Sayers EW, Cavanaugh M, Clark K *et al.* GenBank. *Nucleic Acids Res* 2021;49:D92–6. https://doi.org/10.1093/nar/gkaa1023
24. Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114
25. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* 2016;32:1323–30. https://doi.org/10.1093/bioinformatics/btw006
26. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153
27. Hyatt D, Chen GL, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 2010;11:119. https://doi.org/10.1186/1471-2105-11-119
28. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 2004;32:11–6. https://doi.org/10.1093/nar/gkh152
29. Palmer JM, Stajich J. Funannotate v1. 8.1: Eukaryotic genome annotation. 2020. https://doi.org/10.5281/zenodo.1134477
30. Stanke M, Keller O, Gunduz I *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006;34:W435–9. https://doi.org/10.1093/nar/gkl200
31. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 2007;35:3823–35. https://doi.org/10.1093/nar/gkm238
32. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 2004;20:2878–9. https://doi.org/10.1093/bioinformatics/bth315
33. Testa AC, Hane JK, Ellwood SR *et al.* CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 2015;16:170. https://doi.org/10.1186/s12864-015-1344-4
34. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO *et al.* Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 2005;33:6494–506. https://doi.org/10.1093/nar/gki937
35. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO *et al.* Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res* 2008;18:1979–90. https://doi.org/10.1101/gr.081612.108
36. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* 2014;42:e119. https://doi.org/10.1093/nar/gku557
37. Chan PP, Lin BY, Mak AJ *et al.* tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 2021;49:9077–96. https://doi.org/10.1093/nar/gkab688
38. Russel J, Pinilla-Redondo R, Mayo-Muñoz D *et al.* CRISPRCasTyper: automated identification, annotation, and classification of CRISPR–Cas loci. *CRISPR J* 2020;3:462–9. https://doi.org/10.1089/crispr.2020.0059
39. Yi H, Huang L, Yang B *et al.* AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. *Nucleic Acids Res* 2020;48:W358–65. https://doi.org/10.1093/nar/gkaa351
40. Nielsen H, Teufel F, Brunak S *et al.* SignalP: the evolution of a web server. In: *Protein Bioinformatics*. New York: Springer, 2024, 331–67. https://doi.org/10.1007/978-1-0716-4007-4_17
41. Krogh A, Larsson B, Von Heijne G *et al.* Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305:567–80. https://doi.org/10.1006/jmbi.2000.4315
42. 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
43. 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
44. Consortium GO. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 2017;45:D331–8. https://doi.org/10.1093/nar/gkw1108
45. Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092
46. Schellenberger J, Park JO, Conrad TM *et al.* BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinform* 2010;11:213. https://doi.org/10.1186/1471-2105-11-213
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. 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
49. Buchfink B, Reuter K, Drost HG. 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
50. 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
51. Lu T, Yao B, Zhang C. DFVF: database of fungal virulence factors. *Database* 2012;2012:bas032. https://doi.org/10.1093/database/bas032

52. 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
53. 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
54. Blin K, Shaw S, Augustijn HE *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* 2023;51:W46–50. https://doi.org/10.1093/nar/gkad344
55. Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
56. Zielezinski A, Vinga S, Almeida J *et al.* Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol* 2017;18:186. https://doi.org/10.1186/s13059-017-1319-7
57. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* 2024;52:W78–82. https://doi.org/10.1093/nar/gkae268
58. Wang RH, Yang S, Liu Z *et al.* PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic Acids Res* 2024;52:D756–61. https://doi.org/10.1093/nar/gkad979
59. Li Y, Feng X, Chen X *et al.* PlasmidScope: a comprehensive plasmid database with rich annotations and online analytical tools. *Nucleic Acids Res* 2025;53:D179–88. https://doi.org/10.1093/nar/gkae930
60. Jia W, Li H, Li S *et al.* Oviz-Bio: a web-based platform for interactive cancer genomics data visualization. *Nucleic Acids Res* 2020;48:W415–26. https://doi.org/10.1093/nar/gkaa371
61. 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
62. 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
63. Zhu J, Tian L, Chen P *et al.* Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa. *Genom Proteom Bioinform* 2022;20:246–59. https://doi.org/10.1016/j.gpb.2021.05.001
64. Shaiber A, Willis AD, Delmont TO *et al.* Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* 2020;21:292. https://doi.org/10.1186/s13059-020-02195-w
65. Li Y, Pan G, Wang S *et al.* Comprehensive human respiratory genome catalogue underlies the high resolution and precision of the respiratory microbiome. *Brief Bioinform* 2025;26:bbae620. https://doi.org/10.1093/bib/bbae620
66. Joglekar P, Conlan S, Lee-Lin SQ *et al.* Integrated genomic and functional analyses of human skin–associated Staphylococcus reveal extensive inter-and intra-species diversity. *Proc Natl Acad Sci* 2023;120:e2310585120. https://doi.org/10.1073/pnas.2310585120
67. Huang L, Guo R, Li S *et al.* A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. *Nat Microbiol* 2024;9:2185–200. https://doi.org/10.1038/s41564-024-01751-5
68. Shen H, Wang T, Dong W *et al.* Metagenome-assembled genome reveals species and functional composition of Jianghan chicken gut microbiota and isolation of Pediococcus acidilactic with probiotic properties. *Microbiome* 2024;12:25. https://doi.org/10.1186/s40168-023-01745-1
69. Hu H, Huang Y, Yang F *et al.* Metagenome-assembled microbial genomes ($n = 3,448$) of the oral microbiomes of Tibetan and Duroc pigs. *Sci Data* 2025;12:141. https://doi.org/10.1038/s41597-025-04413-1
70. Xu S, Huang H, Chen S *et al.* Recovery of 1887 metagenome-assembled genomes from the South China Sea. *Sci Data* 2024;11:197. https://doi.org/10.1038/s41597-024-03050-4
71. Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;49:D412–419. https://doi.org/10.1093/nar/gkaa913
72. Terlouw BR, Blin K, Navarro-Munoz JC *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res* 2023;51:D603–10. https://doi.org/10.1093/nar/gkac1049
73. Uchiyama I. MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res* 2003;31:58–62. https://doi.org/10.1093/nar/gkg109
74. Davis JJ, Wattam AR, Aziz RK *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res* 2020;48:D606–12.
75. Davidsen T, Beck E, Ganapathy A *et al.* The comprehensive microbial resource. *Nucleic Acids Res* 2010;38:D340–5. https://doi.org/10.1093/nar/gkp912

Received: August 14, 2025. **Revised:** October 16, 2025. **Accepted:** October 16, 2025