

PlasmidScope: a comprehensive plasmid database with rich annotations and online analytical tools

Yinhu Li^{1,3,8,†}, Xikang Feng^{2,4,†}, Xuhua Chen^{3,†}, Shuo Yang⁵, Zicheng Zhao⁷, Yu Chen^{1,3,8,*} and Shuai Cheng Li^{5,6,*}

¹CAS Key Laboratory of Brain Connectome and Manipulation, the Brain Cognition and Brain Disease Institute, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

²School of Software, Northwestern Polytechnical University, Xi'an 710072, China

³SIAT-HKUST Joint Laboratory for Brain Science, Chinese Academy of Sciences, Shenzhen 518055, China

⁴Research & Development Institute, Northwestern Polytechnical University, Shenzhen 518063, China

⁵Department of Computer Science, City University of Hong Kong, Hong Kong, China

⁶City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China

⁷OmicLab Limited, Science Park East Avenue, Hong Kong Science Park, Hong Kong, China

⁸Shenzhen-Hong Kong Institute of Brain Science, Shenzhen Fundamental Research Institutions, Shenzhen 518055, China

*To whom correspondence should be addressed. Tel: +852 3442 9412, Fax: +852 3442 0503, Email: shuaicli@cityu.edu.hk

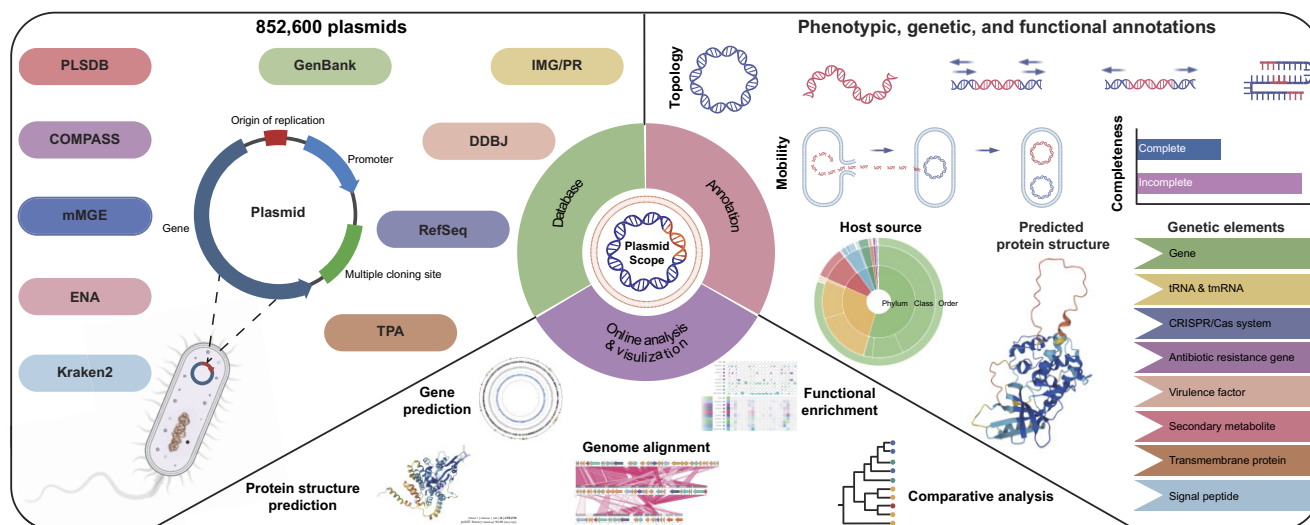
Correspondence may also be addressed to Yu Chen. Tel: +86 755 26925498, Fax: +86 755 26925498, Email: yu.chen@siat.ac.cn

[†]The first three authors should be regarded as Joint First Authors.

Abstract

Plasmids are extrachromosomal genetic molecules that replicate independent of chromosomes in bacteria, archaea, and eukaryotic organisms. They contain diverse functional elements and are capable of horizontal gene transfer among hosts. While existing plasmid databases have archived plasmid sequences isolated from individual microorganisms or natural environments, there is a need for a comprehensive, standardized, and annotated plasmid database to address the vast accumulation of plasmid sequences. Here, we propose PlasmidScope (<https://plasmid.deepomics.org/>), a plasmid database offering comprehensive annotations, automated online analysis, and interactive visualization. PlasmidScope harbors a substantial collection of 852 600 plasmids curated from 10 repositories. Along with consolidated background information, PlasmidScope utilizes 12 state-of-the-art tools and provides comprehensive annotations for the curated plasmids, covering genome completeness, topological structure, mobility, host source, tRNA, tmRNA, signal peptides, transmembrane proteins and CRISPR/Cas systems. PlasmidScope offers diverse functional annotations for its 25 231 059 predicted genes from 9 databases as well as corresponding protein structures predicted by ESMFold. In addition, PlasmidScope integrates online analytical modules and interactive visualization, empowering researchers to delve into the complexities of plasmids.

Graphical abstract



Received: August 14, 2024. Revised: October 2, 2024. Editorial Decision: October 3, 2024. Accepted: October 8, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

A plasmid is an extrachromosomal genetic molecule that replicates independent of chromosomes in bacteria, archaea, and eukaryotic organisms. They carry diverse functional elements and facilitates horizontal gene transfer (1,2). As a class of mobile genetic elements, plasmids have diverse functional elements, such as antibiotic resistance genes (ARGs) (3), virulence factors (VFs) (4), and metabolic compounds (5). They also contribute to genetic diversity (6,7) and rapid adaptation of the host to various environments (8). For example, ARGs in plasmids enable the survival of pathogens under antibiotic exposure, leading to antibiotic-resistant infections, which is a serious challenge in clinical settings (8–10). In addition, the horizontal gene transfer capabilities of plasmids, including conjugation, transformation, and transduction, enhance the spread of genetic elements across different microorganisms and environments (11), which might accelerate host evolution by providing highly diverse and multicopy plasmid-encoded genes. Therefore, plasmids play crucial roles in evolutionary dynamics and ecological diversity.

The physiological features of plasmids, such as topology and mobility, as well as their genetic elements, including genes, signal peptides, and CRISPR/Cas systems, provide a foundation for understanding horizontal gene transfer, evolutionary dynamics, and ecological diversity (12–16). Topology and mobility are key features of the horizontal gene transfer capability of plasmids: topology influences plasmid stability and replication within hosts (12,17), while mobility determines whether plasmids can spread among different microorganisms through conjugation, transformation, or transduction (12,13). Plasmid-encoded genes are the basic genetic materials spread in different microbial populations and promote evolutionary dynamics (14). Plasmid-encoded signal peptides direct proteins to the cell membrane or secretion systems, facilitating signal transmission and promoting ecological diversity (18). In addition, plasmid-carried CRISPR/Cas systems enable the host to acquire immune memory, allowing the quick adaption of the hosts to viral or foreign DNA threats (19). Thus, the physiological and genetic features of plasmids are crucial for understanding plasmid structures and have bioengineering or clinical applications.

Several plasmid databases that organize the physiological and genetic features of plasmids have been developed in recent years. Plasmid databases, such as PLSDb (20), COMPASS (21) and pATLAS (22), retrieved reliable plasmids from GenBank or the RefSeq database (23,24), providing genetic characteristics and host information as well as supporting web-based data manipulation to assist plasmid-related queries and identification. However, these plasmid databases are typically limited to single microorganisms and therefore do not capture the diversity of plasmids in natural environments (25). With the immense accumulation of next-generation sequencing data and the rapid development of plasmid identification algorithms, some plasmid databases, such as IMG/PR (25) and mMGE (26), have expanded the identification of plasmids based on assembled metagenomes and metatranscriptomes, boosting the number of cataloged plasmids from tens of thousands to millions. Yet, these plasmid databases seldom provide information about the structures of proteins or genetic elements, such as signal peptides, transmembrane proteins, and CRISPR/Cas systems, nor do they facilitate cross-comparison or querying among different databases.

To this end, we propose PlasmidScope (<https://plasmid.deepomics.org/>), a comprehensive plasmid database offering rich annotations, interactive visualization, and online analysis of curated plasmids. PlasmidScope contains an extensive collection of over 1 million plasmids curated from 10 public repositories. PlasmidScope also provides detailed, downloadable annotations for its curated plasmids, including topological annotation, mobility prediction, completeness assessment, host taxonomy, functional annotations, protein structure prediction, signal peptide and transmembrane protein prediction and CRISPR/Cas system prediction. In addition, PlasmidScope supports interactive visualization of its curated database and related annotation results and contains online analytical modules for custom plasmid analysis.

Materials and methods

Plasmid collection and curation

To amass a comprehensive collection of plasmid sequences, we performed an extensive search across public repositories, including GenBank (23), RefSeq (24), ENA (27), DDBJ (28), Kraken2 (29) and TPA (30), focusing on the keyword ‘plasmid’. We searched ‘plasmid’ under the Nucleotide category on the NCBI website. Given that the GenBank, ENA and DDBJ databases are synchronized through the INSDC initiative, we selected the options for ‘GenBank’, ‘DDBJ’, ‘ENA’, ‘RefSeq’, ‘RefSeq’, ‘PDB’ and ‘TPA’ under the ‘Source Databases’ section on the search page. This allowed us to download plasmids while differentiating by their source database. We also included plasmids gleaned from published datasets, such as PLSDb (20), COMPASS (21), IMG/PR (25) and mMGE (26) (Supplementary Table S1). Within these datasets, our collection extended beyond plasmid sequences to include host taxonomic data, topological structures, and assessment of their completeness.

Following plasmid data collection, we curated the data as follows: First, we filtered out plasmids with a genomic length less than 200 bp, GC content less than 10%, or GC content >90%. Second, we removed the duplicate plasmids with identical plasmid sequences within each database. Third, we identified the non-redundant plasmids across the 10 plasmid databases (Supplementary Figure S1). To identify the duplicated plasmids, we applied MMseqs2 (version 15.6f452) to detect plasmids with identical sequences (clustering parameters: ‘-cov-mode 0 -c 1.0 -min-seq-id 1.0’) (31). We retained the deduplicated plasmids as the final curated plasmid database and included them in the “All” table in the ‘Plasmid list’ page on the website of PlasmidScope. For the redundant plasmids with multiple entries, we consolidated them into the same row of the “All” table and provided links to their source databases. To enhance the connectivity between PlasmidScope and the source databases, we also provide other tables tagged with ‘PLSDb’, ‘IMG/PR’, ‘COMPASS’, ‘ENA’, ‘GenBank’, etc., containing original plasmids from the specified databases.

Host taxonomic assignment

We collected the host taxonomic data from each dataset. Given that the above datasets were released at different times and that the taxonomic information might vary by version, we performed rigorous standardization. Accordingly, we aligned the host taxonomic information across all datasets to con-

form to the most current version available from the NCBI (i.e. 202406) (32). This ensures a unified and up-to-date reference point for the host species associated with each plasmid, facilitating the accuracy and consistency of information.

Plasmid clustering and mobility prediction

To identify plasmid cluster assignments, we used the MOB-typer tool from MOB-suite (v3.1.8) (33) using the default parameters and databases. To obtain higher clustering resolution, we provided cluster and subcluster information for the plasmids after MOB-typing analysis. We applied Mash embedded in MOB-suite for plasmid clustering based on the fast genomic distance estimation. This strategy leverages the genomic sequence of the plasmids for analysis, transcending the limitations of relying on specific biomarkers. MOB-typer also predicts the mobility of plasmids as conjugative, mobilizable, or non-mobilizable (33).

Annotation of plasmid genes and genomic elements

We performed gene annotation for the plasmid genomes using Prokka (v1.11) (34), which integrates Prodigal (v2.6.3) (35), to identify the open reading frames, along with ARAGORN (v1.2.41) (36), which detects tRNA and tmRNA genes in the plasmid sequences. We subsequently analyzed CRISPR/Cas systems within the curated plasmids. To detect the CRISPR arrays and associated Cas genes in each plasmid genome, we used CRISPRCasTyper (v1.8.0) (37). Furthermore, we identified the subtypes of these systems based on a comprehensive assessment of both the Cas genes and the CRISPR repeat sequences. We subsequently employed SignalP 6.0 (38) to predict the presence of signal peptides and the location of their cleavage sites in proteins. Finally, to detect the topology of membrane proteins, we used TMHMM 2.0 (39), which utilizes a hidden Markov model to identify the structural intricacies of membrane proteins. We used all of the abovementioned tools with their respective default parameters (Supplementary Table S2).

Functional annotation

For functional annotation of coding sequences, we applied eggNOG-mapper (v2.1.12) (40) with the default parameters to the fast orthology assignments using precomputed eggNOG (v5.0.2) (41) clusters and phylogenies. The results of eggNOG annotation included a range of matching and scoring information as well as functional annotation insights from various databases, such as Gene Ontology (GO) (42), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (43), the BiGG Database (44), Clusters of Orthologous Groups (COG) (45) and the Carbohydrate-Active EnZymes database (CAZy) (46) (Supplementary Table S3). Furthermore, we employed Diamond (v2.1.8.162) (47) to conduct a homology search for plasmid proteins against the Virulence Factor Database (VFDB) (48), identifying virulence factors when matches exceeded a 60% threshold of identity and a 40% threshold of coverage. We performed antibiotic resistance gene annotation based on both homology and single-nucleotide polymorphism models using reference data from the Comprehensive Antibiotic Resistance Database (CARD) (49) with the following parameters: `'-include_loose'` and `'-include_nudge'`. Finally, we employed antiSMASH (v7.1.0) (50) to identify and annotate the secondary metabolite biosynthesis gene clusters

within the plasmid genomes using the following parameters: `'-asf -cc-mibig -cb-general -cb-knownclusters -cb-subclusters -pfam2go'`.

Protein structure prediction

We predicted the sophisticated structures of the proteins and generated 3D models using advanced artificial intelligence-based modeling methods. Leveraging the large language model, ESMFold predicts full atomic-level protein structures directly from the primary sequence (51). Recognizing the potential of this advancement, we have embedded ESMFold into the web-based interface of PlasmidScope. For each predicted protein structure, we adopted the predicted local distance difference test (pLDDT) to assess the prediction quality for its residues. On the 'Protein detail' page, users can hover over the structure to view the pLDDT value for each residue and download the CIF file that contains pLDDT values for further application.

Sequence alignment

To compare coding DNA sequences across a spectrum of plasmid genomes provided by users or against the plasmid genomes in PlasmidScope, we implemented BLASTP (52) to facilitate pairwise alignment of the encoded proteins. In the alignment visualizations, we showcase the alignment coverage and identity values derived from the BLAST outputs. The plasmid order in these visualizations is automatically determined to ensure optimal alignment arrangement, indicating closer genetic resemblance within each group.

Comparative analysis

We constructed a comparative tree that illustrates the relationship among multiple plasmids using a 2-step process. First, we used Alfpay (53), an alignment-free sequence comparison method, to calculate the genomic distance (Euclidean distance based on k -mer = 6) between the plasmid sequences. We subsequently applied the neighbor-joining algorithm to synthesize these distances into a comparative tree, thereby providing a visual representation of the relationships among the plasmids.

Statistical analysis

In the case study, we analyzed the length distribution of the plasmids using the ggplot2 and plyr packages in R (54,55). Based on the functional annotations obtained from PlasmidScope, we calculated the number of genes in each COG, ARG and VF category and visualized the enrichment results for the conjugative, mobilizable, and non-mobilizable plasmids using the sankeywheel package in R (<https://cran.r-project.org/web/packages/sankeywheel/vignettes/sankeywheel.html>). In addition, we calculated the gene numbers for the ARG categories within each plasmid and compared the distribution of cephalosporin resistance genes across the conjugative, mobilizable, and non-mobilizable plasmids using the Wilcoxon rank-sum test. The level of statistical significance was set at $P < 0.05$.

Platform development

PlasmidScope 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 Apache, Django, PostgreSQL and

Typescript+Vue3 (56,57). All online data visualizations are implemented using Oviz (58). We also provide detailed tutorials on the platform to facilitate user navigation and utilization.

Results

Comprehensive annotated plasmid database in PlasmidScope

PlasmidScope contains a comprehensive database of 852 600 plasmids curated from 10 repositories (Figure 1), including 8 repositories that isolate plasmids from single microorganisms and 2 repositories that incorporate metagenomic and metatranscriptomic data (Supplementary Table S1). The plasmid repositories, PLSDb (20), COMPASS (21), GenBank (23), RefSeq (24), ENA (27), DDBJ (28), Kraken2 (29) and TPA (30), contain 110,021 plasmids isolated from single microorganisms, constituting 12.90% of the PlasmidScope database. Meanwhile, the plasmid repositories that incorporate metagenomic and metatranscriptomic data, IMG/PR (25) and mMGE (26), adopt homologous sequence alignment or deep learning to capture 742,579 plasmids from assembled metagenomes and metatranscriptomes, constituting 87.10% of the PlasmidScope database. To broaden the applications of these plasmids, PlasmidScope systematically encompasses each curated plasmid with standardized and comprehensive annotations that can be summarized into 3 main categories: feature, genetic feature and gene feature.

PlasmidScope collects and consolidates six features of the curated plasmids: genomic length, GC content, genome completeness, topological structure, mobility and host source (Figure 1). The genomic length of the plasmids ranges from 200 to 11 850 240 bp, while their GC content ranges from 11.73% to 87.48%. Among the curated plasmids, 208 350 have complete genomes. Regarding topological structure, PlasmidScope categorizes plasmids as circular plasmids (75 454 sequences), linear plasmids (645 375 sequences), direct terminal repeats (124 254 sequences), inverted terminal repeats (2577 sequences), or concatemers (4940 sequences). Meanwhile, 696 657 (81.71%) plasmids are non-mobilizable plasmids, followed by 106 800 (12.53%) mobilizable and 49 143 (5.76%) conjugative plasmids. In addition, the host distributions of 360 583 (42.29%) plasmids are available in PlasmidScope, which are mainly abundant in Enterobacterales (28.63%), Bacillales (9.24%) and Bacteroidales (7.92%). These precisely integrated features assist users in exploring the basic characteristics of plasmids and selecting appropriate and efficient plasmids for further use, such as bioengineering vector selection.

PlasmidScope provides six genetic features of the curated plasmids: gene, tRNA, tmRNA, signal peptide, transmembrane protein and CRISPR/Cas system (Figure 1, Supplementary Table S4). By screening the plasmids through the standardized bioinformatic pipeline, we identified 25 231 059 genes, 81 963 tRNAs and 755 tmRNAs. In addition, PlasmidScope includes 2,710,395 signal peptides that can be categorized into five classes: Sec signal peptide (SP, 71.43%), Lipoprotein signal peptide (LIPO, 22.00%), Tat signal peptide (TAT, 5.06%), Tat lipoprotein signal peptide (TATLIPO, 0.50%), and Pilin signal peptide (PILIN, 1.01%). PlasmidScope provides detailed structural information for 5 191 488 identified transmembrane proteins, including the number, order and amino acid sequences of the transmembrane seg-

ments, thereby facilitating the exploration of the effects of structures on the transport of specific substances. Furthermore, PlasmidScope offers detailed information on 4083 predicted CRISPR/Cas systems, including types, subtypes, Cas genes and consensus repeat sequences. The exhaustive annotated genetic features enable users to investigate the structures of genetic elements, facilitating related structure design and optimization.

PlasmidScope contains comprehensive functional annotations and advanced translated protein structural prediction for each predicted gene (Figure 1). To investigate the functions of plasmids, PlasmidScope performs functional annotations for the 25,231,059 predicted genes with 9 databases, including GO (42), KEGG (43), BiGG (44), COG (45), CAZy (46), VFDB (48), CARD (49), Pfam (59) and MIBiG (60). In addition, PlasmidScope provides the specific functional category information underlying each database, such as COG category, GO class and KEGG pathway. To provide an in-depth view into the natural diversity of proteins, PlasmidScope predicts the protein structures of genes with ESMFold and produces high-resolution visualizations of protein structures that show atomic-level structures. Notably, PlasmidScope supports both online viewing adjustment of protein structures and the downloading of protein structure files. The comprehensive functional annotations and advanced protein structure predictions can accelerate the discovery of horizontal-transferring functional genes and the detection of protein structures.

Integrated online analytical modules in PlasmidScope

PlasmidScope offers online analytical modules that allow users to upload their custom plasmids, perform functional and structural annotations, and make comparisons with plasmids in the database. Users can upload single or multiple plasmid sequences and choose from 6 analytical modules: open reading frame prediction and protein annotation, tRNA and tmRNA prediction, ARG and VF annotation, transmembrane protein annotation, sequence alignment, and comparative analysis—all of which utilize the advanced tools detailed in the Materials and methods section (Figure 2). Furthermore, PlasmidScope integrates these six modules into two analytical pipelines according to their application field: plasmid annotation and plasmid comparison. The plasmid annotation pipeline includes open reading frame prediction and protein annotation, tRNA and tmRNA prediction, ARG and VF annotation, and transmembrane protein annotation. Meanwhile, the plasmid comparison pipeline includes sequence alignment and comparative analysis. Users can perform either single or multiple analyses based on their specific research needs and objectives. After the analysis is complete, users can check and download the resulting documents and visualizations, supporting further plasmid research.

Informative and interactive visualizations in PlasmidScope

PlasmidScope features a user-friendly interface that enables users to seamlessly query, visualize, analyze, and download information. PlasmidScope has 7 menus: 'Home', 'Database', 'Analysis', 'Workspace', 'Download', 'Tutorial' and 'Contact Us'. The homepage provides a basic summary of PlasmidScope, three key feature plots, and version updates. The 'Database' menu offers a concise summary page (Figure 3),

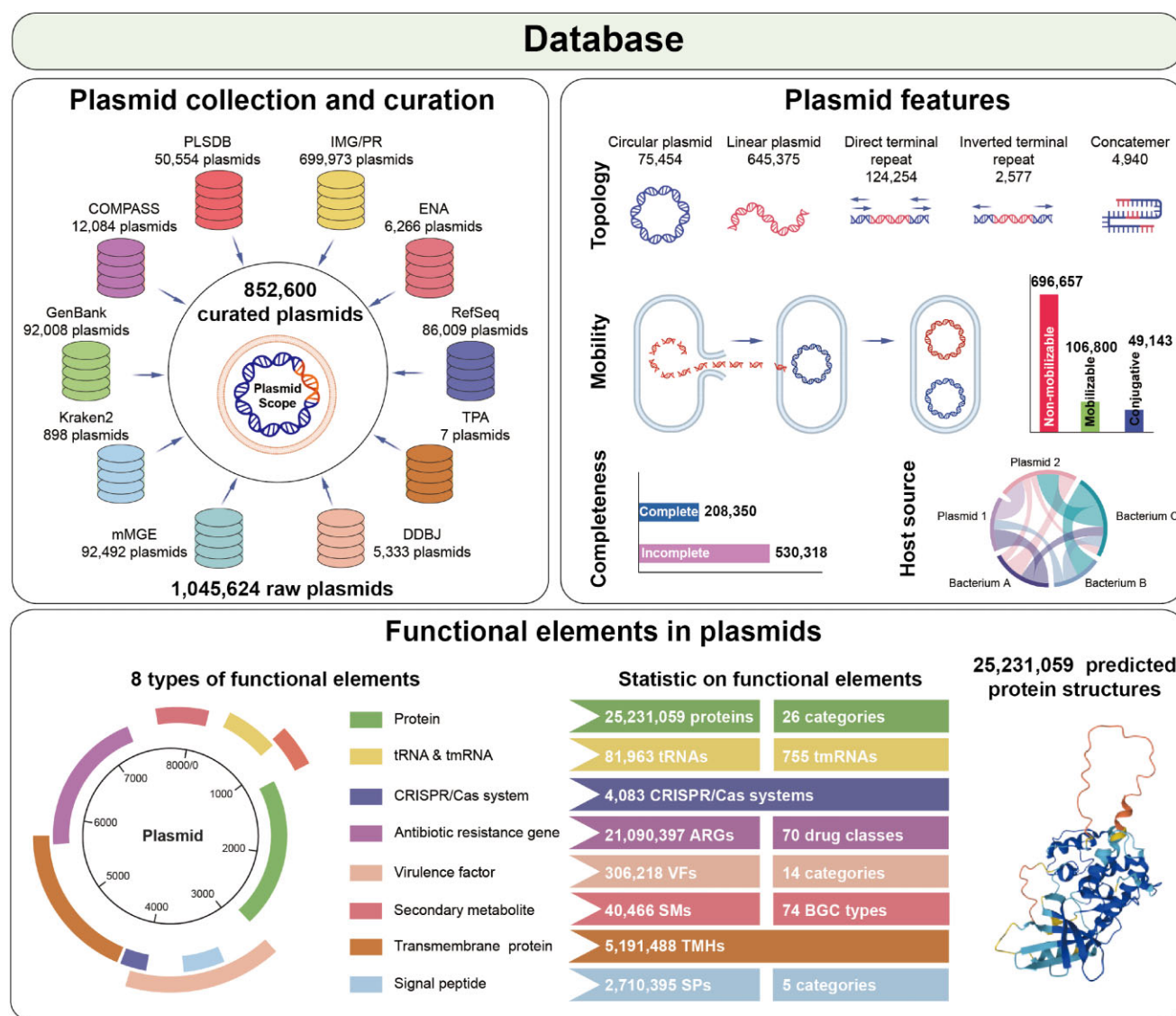


Figure 1. Overview of the plasmid database in PlasmidScope. PlasmidScope contains a comprehensive database of 852,600 plasmids curated from 10 repositories along with standardized and comprehensive annotations, including feature consolidation, genetic feature annotation, gene function annotation, and protein structure prediction.

11 detailed annotation pages, and a querying page. Each of the 11 annotation pages presents key features from various perspectives, such as plasmid feature, host source, cluster information, and protein annotation, along with a search box in the upper-right corner. In addition, some annotation pages contain 'Detail' and 'Download' buttons, allowing users to view detailed annotation results and download related sequences, respectively. As described above, the 'Analysis' menu comprises two pages: 'Plasmid Annotation' and 'Plasmid Comparison', which contain four and two analytical modules, respectively. The 'Workspace' menu displays the status of submitted tasks. Users can click the 'View Result' button for each task to access and download their custom analysis and visualization results, which can be easily integrated into academic publications (Figure 3). The 'Download' menu provides a list of available files for download. Finally, the 'Tutorial' menu offers user detailed instructions on how to use PlasmidScope, and the 'Contact Us' menu facilitates convenient communication with the support team.

Case study: plasmids as reservoirs of antibiotic-resistance genes in *Klebsiella pneumoniae*

As an opportunistic pathogen, *Klebsiella pneumoniae* can cause pneumonia, urinary tract infection, bacteremia, and other infections when the host's immune function is compromised (61). In recent years, the overuse of antibiotics has led to the emergence and global spread of hypervirulent *K. pneumoniae*, which has been reported in 43 countries and exhibits resistance to third-generation cephalosporins, posing a significant threat to public health (<https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON527>). As one of the main vehicles for horizontal gene transfer, plasmids offer opportunities to deepen our understanding on the emergence of *K. pneumoniae* multidrug resistance.

Accordingly, we used PlasmidScope to analyze the distribution and characteristics of ARGs in plasmids from *K. pneumoniae*. By accessing the plasmids under the 'Host list' page in PlasmidScope, we collected 21 007 plasmids from

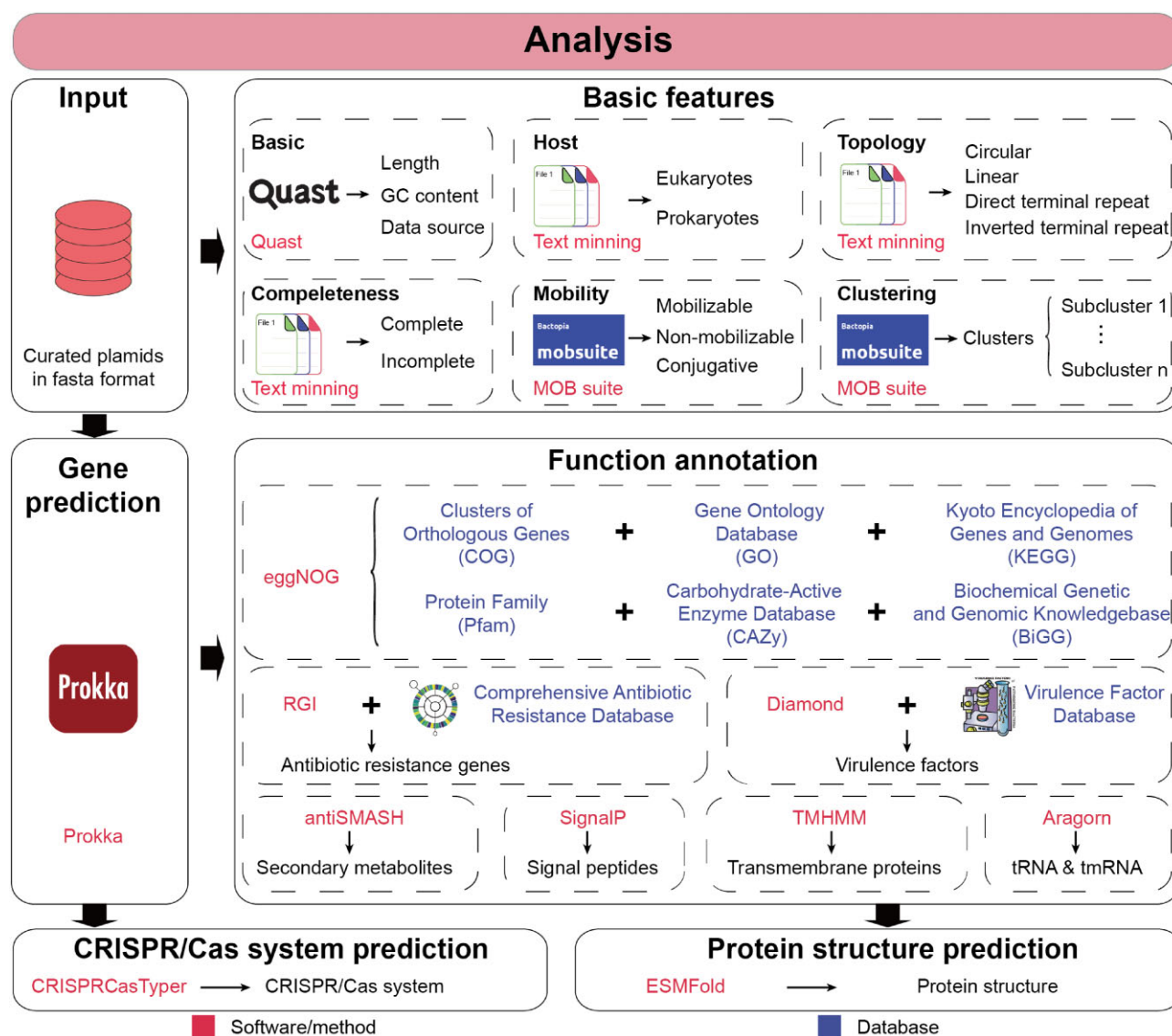


Figure 2. Overview of the analytical procedure of PlasmidScope. PPlasmidScope uses bioinformatics tools and eight functional databases to perform standardized annotations for each plasmid sequence. In addition, PlasmidScope contains six analytical modules and two analytical pipelines, enabling users to perform online plasmid analysis.

K. pneumoniae, including 7010 conjugative, 4132 mobilizable and 9865 non-mobilizable plasmids (Figure 4A). Using the feature annotations provided by PlasmidScope, we further identified 12 175 complete plasmids and 11 212 circular plasmids (Figure 4A). Upon analyzing plasmid length distributions, we found that the conjugative plasmids had the longest genome ($127\,075.43 \pm 83\,573.62$ bp), followed by the non-mobilizable ($37\,986.33 \pm 83\,788.18$ bp) and mobilizable plasmids ($28\,704.41 \pm 42\,249.88$ bp) (Figure 4B). These plasmids harbored 1 579 934 predicted genes, including 1 234 887 ARGs and 15 575 VFs (Figure 4C–E). The conjugative, mobilizable and non-mobilizable plasmids contained substantial COGs particularly enriched in the categories of ‘replication, recombination and repair’ (17.00%, 22.21% and 17.40%, respectively), ‘transcription’ (4.46%, 5.59% and 5.72%, respectively), and ‘inorganic ion transport and metabolism’ (4.03%, 2.83% and 5.33%, respectively) (Figure 4C). ARG enrichment indicated that all three types of plasmids mainly had resistance against ‘penam’ (11.50%) and

‘cephalosporin’ (9.44%) (Figure 4D). VF enrichment analysis revealed that the ‘invasion’ function was abundant in the conjugative and mobilizable plasmids (24.47% and 28.82%, respectively), but less prevalent in the non-mobilizable plasmids (9.99%) (Figure 4E). Furthermore, the conjugative and mobilizable plasmids carried significantly more antibiotic resistance genes against cephalosporin than the non-mobilizable plasmids ($P < 0.001$, Figure 4F), offering insights into the acquisition of multidrug-resistance in *K. pneumoniae*. Thus, these findings demonstrate the utility of PlasmidScope as a convenient tool in the face of the antimicrobial resistance crisis.

Discussion

To our knowledge, PlasmidScope is the largest and most comprehensive plasmid database that integrates built-in analytical tools and interactive visualization to help researchers decipher the physiological and genetic characteristics of plasmids

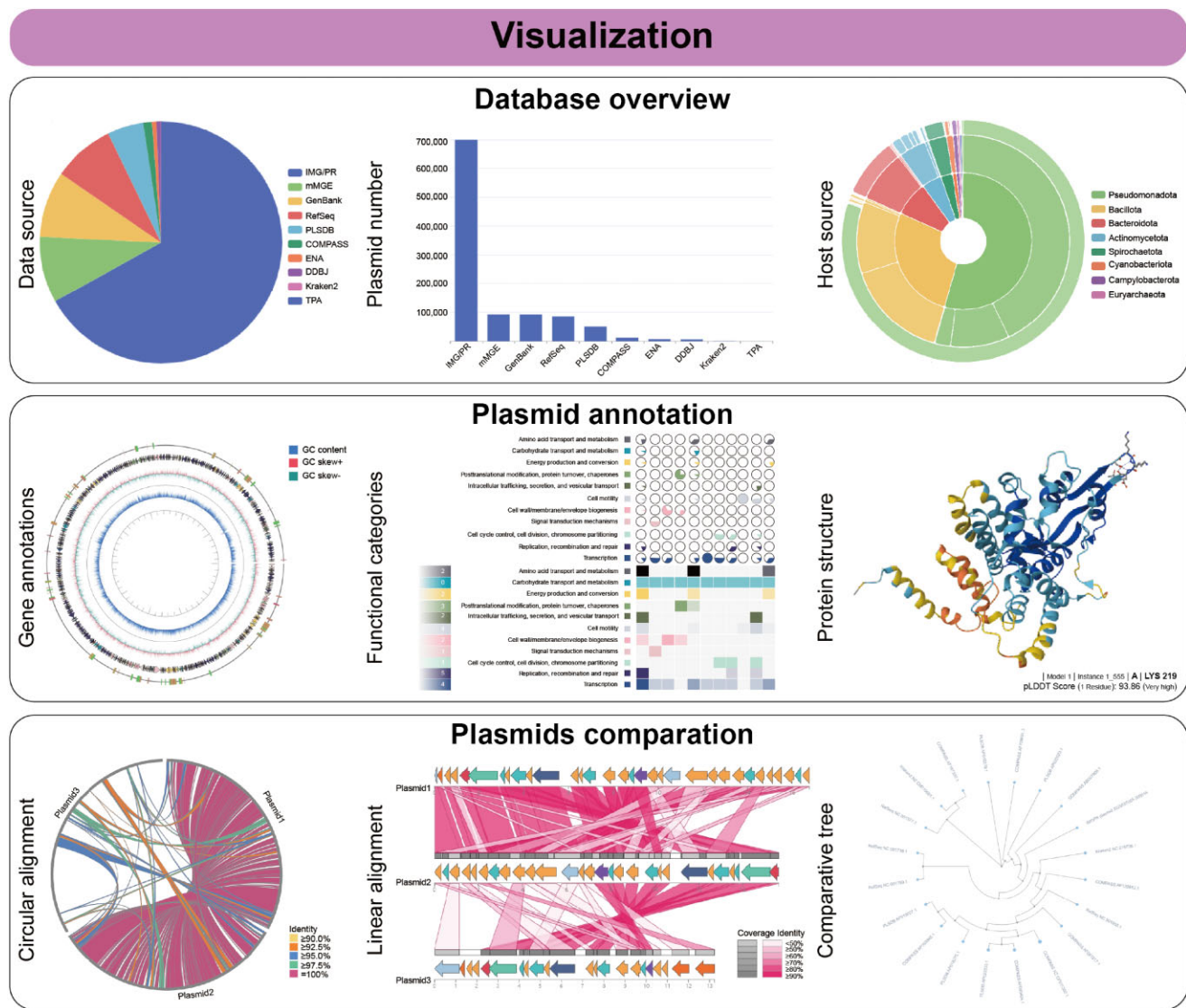


Figure 3. Overview of interactive visualizations in PlasmidScope. PlasmidScope generates interactive visualizations of its curated database along with comprehensive annotations. In addition, analysis results and visualizations generated from the online analytical modules or pipelines of PlasmidScope are downloadable.

across different microbial hosts. PlasmidScope has the following key features: (i) an extensive collection of over 1 million plasmids with standardized and comprehensive annotations; (ii) online analytical modules that support customized plasmid annotation and plasmid comparison; (iii) interactive visualization of the curated database, detailed annotations, and customizable analysis results; (iv) downloadable support for all resources. PlasmidScope is freely available to the public and can be accessed without registration.

In addition to its extensive collection of plasmids, PlasmidScope has several key advantages. First, PlasmidScope is closely integrated with existing databases, retaining the IDs and background information of plasmids in other databases, which enables connections and cross-comparison among databases. Second, PlasmidScope enables comprehensive, accurate, and downloadable annotations, which are essential for investigating plasmid genetic features. Third, PlasmidScope contains integrated online analytical modules to support plasmid annotation, plasmid comparison, and visualization. Fourth, PlasmidScope offers informative and interac-

tive visualizations for the curated database, related annotations, and online analysis results. Finally, PlasmidScope integrates ESMFold to help users predict and view protein structures, which is not possible with other plasmid databases. Hence, PlasmidScope is a powerful platform that supports plasmid genome research and genome-scale genetic element analysis.

By serving the scientific community, PlasmidScope can become a centralized platform for plasmid research. We will also continually update and enhance PlasmidScope. First, we plan to expand PlasmidScope by incorporating additional plasmid databases to broaden its applicability and facilitate cross-referencing of information across studies. Second, we plan to embed more deep-learning models and bioinformatic modules, such as completeness assessment and host prediction, to support online analysis and data mining. Third, we will continuously optimize PlasmidScope's framework and interface based on user feedback, enhancing its usability. Last, we aim to establish a secure and reliable system to facilitate big data exchange and support the archiving of plasmid data for users.

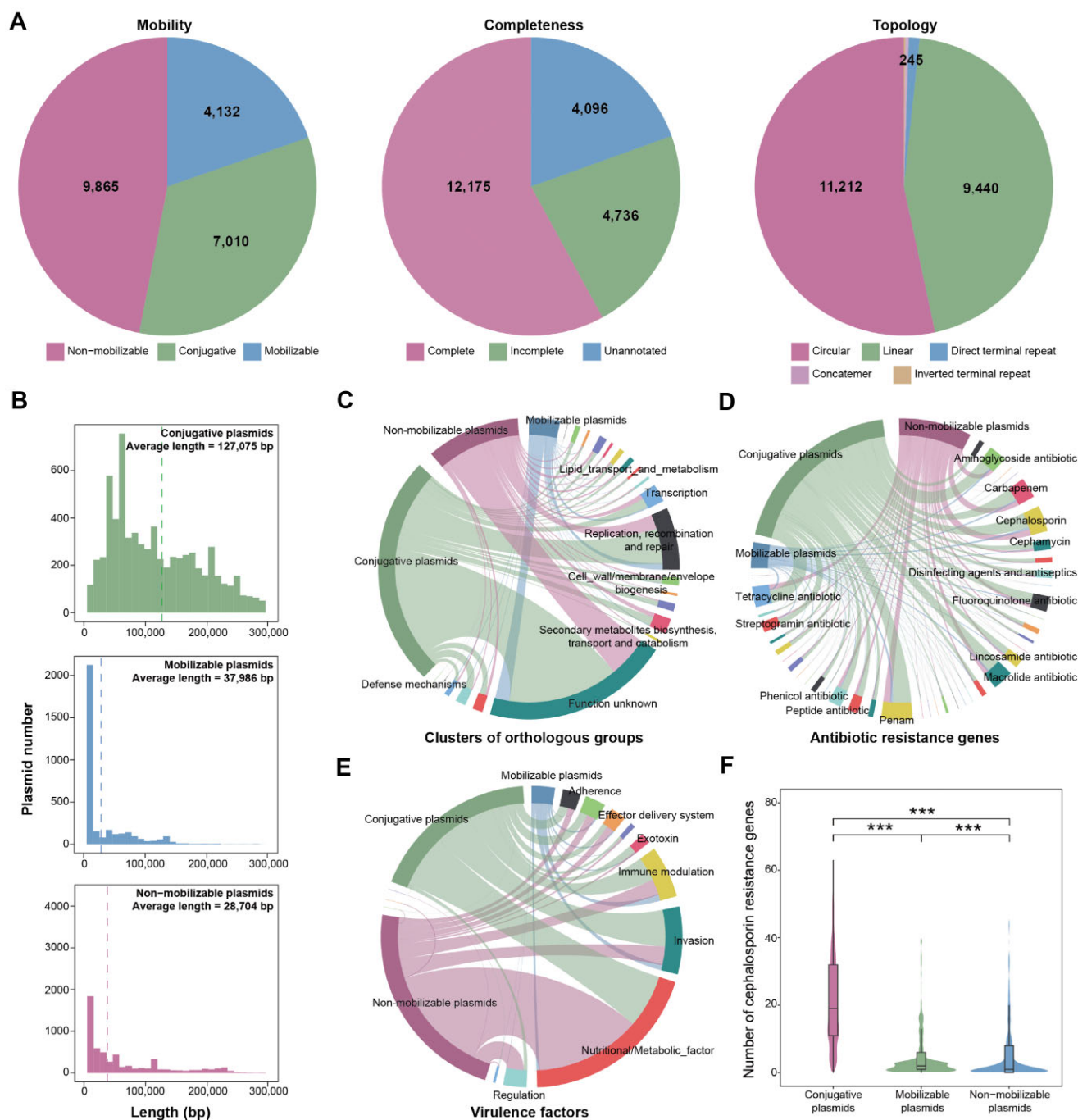


Figure 4. Case study: using PlasmidScope to investigate the antibiotic resistance genes in plasmids from *Klebsiella pneumoniae*. **(A)** Mobility, completeness and topological features in the 21 007 plasmids from *K. pneumoniae*. **(B)** Length distributions of the conjugative, mobilizable and non-mobilizable plasmids. **(C–E)** Gene enrichment results based on the Clusters of Orthologous Groups database **(C)** Comprehensive Antibiotic Resistance Database **(D)** and Virulence Factor Database. **(E)** Dashed lines indicate the average length of the plasmids. **(F)** Cephalosporin resistance gene numbers across the conjugative, mobilizable and non-mobilizable plasmids. *** $P < 0.001$, Wilcoxon rank-sum test.

In conclusion, PlasmidScope is a valuable resource with practical tools that will contribute to productive research endeavors and empower researchers to delve deeper into the intricacies of plasmids.

Data availability

All data are freely available at <https://plasmid.deepomics.org/>.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank the members of the Li Laboratory and Chen Laboratory for their helpful discussions and insights.

Author contributions: Yinhu Li: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing—

original draft. Xikang Feng: Software, Formal analysis, Methodology, Funding acquisition, Writing—original draft. Xuhua Chen: Data curation, Formal analysis, Methodology, Writing—original draft. Shuo Yang: Software, Writing—review & editing. Zicheng Zhao: Data curation, Software. Yu Chen: Conceptualization, Supervision, Funding acquisition, Writing—review & editing. Shuai Cheng Li: Methodology, Conceptualization, Supervision, Funding acquisition, Writing—review & editing.

Funding

NSFC-RGC Joint Research Scheme [32061160472]; National Natural Science Foundation of China [32300527]; Guangdong Basic and Applied Basic Research Foundation [2022A1515110784 and 2023A1515110450]; Shenzhen Science and Technology Program [20220814183301001, JCYJ20220818100800001, ZDSYS20200828154800001]. Funding for open access charge: NSFC-RGC Joint Research Scheme [32061160472]; Guangdong Basic and Applied Basic Research Foundation [2023A1515110450]; Shenzhen Science and Technology Program [20220814183301001, JCYJ20220818100800001, ZDSYS20200828154800001].

Conflict of interest statement

None declared.

References

- Actis, L.A., Tolmasky, M.E. and Crosa, J.H. (1999) Bacterial plasmids: replication of extrachromosomal genetic elements encoding resistance to antimicrobial compounds. *Front. Biosci.*, **4**, D43–D62.
- Thomas, C.M. (2000) Paradigms of plasmid organization. *Mol. Microbiol.*, **37**, 485–491.
- Jacob, A.E. and Hobbs, S.J. (1974) Conjugal transfer of plasmid-borne multiple antibiotic resistance in *Streptococcus faecalis* var. *zymogenes*. *J. Bacteriol.*, **117**, 360–372.
- Lan, R., Stevenson, G. and Reeves, P.R. (2003) Comparison of two major forms of the *Shigella* virulence plasmid pINV: positive selection is a major force driving the divergence. *Infect. Immun.*, **71**, 6298–6306.
- Brinkmann, H., Goker, M., Koblezek, M., Wagner-Dobler, I. and Petersen, J. (2018) Horizontal operon transfer, plasmids, and the evolution of photosynthesis in Rhodobacteraceae. *ISME J.*, **12**, 1994–2010.
- Zhu, S., Hong, J. and Wang, T. (2024) Horizontal gene transfer is predicted to overcome the diversity limit of competing microbial species. *Nat. Commun.*, **15**, 800.
- Klumper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sorensen, S.J. and Smets, B.F. (2015) Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.*, **9**, 934–945.
- San Millan, A. (2018) Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol.*, **26**, 978–985.
- Wein, T., Hultner, N.F., Mizrahi, I. and Dagan, T. (2019) Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nat. Commun.*, **10**, 2595.
- Fursova, N.K., Kislichkina, A.A. and Khokhlova, O.E. (2022) Plasmids carrying antimicrobial resistance genes in Gram-negative bacteria. *Microorganisms*, **10**, 1678.
- Wein, T. and Dagan, T. (2020) Plasmid evolution. *Curr. Biol.*, **30**, R1158–R1163.
- Tran, F. and Boedicker, J.Q. (2019) Plasmid characteristics modulate the propensity of gene exchange in bacterial vesicles. *J. Bacteriol.*, **201**, 1128.
- Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P. and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
- Rodriguez-Beltran, J., DelaFuente, J., Leon-Sampedro, R., MacLean, R.C. and San Millan, A. (2021) Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.*, **19**, 347–359.
- Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.
- Carr, V.R., Shkoporov, A., Hill, C., Mullany, P. and Moyes, D.L. (2021) Probing the mobilome: discoveries in the dynamic microbiome. *Trends Microbiol.*, **29**, 158–170.
- Higgins, N.P. and Vologodskii, A.V. (2015) Topological behavior of plasmid DNA. *Microbiol. Spectr.*, **3**, 1128.
- Freudl, R. (2018) Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.*, **17**, 52.
- Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K. and Semenov, E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.*, **3**, 945.
- Schmartz, G.P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Muller, R. and Keller, A. (2022) PLSDb: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.*, **50**, D273–D278.
- Douarre, P.E., Mallet, L., Radomski, N., Felten, A. and Mistou, M.Y. (2020) Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of IncF plasmids. *Front. Microbiol.*, **11**, 483.
- Jesus, T.F., Ribeiro-Goncalves, B., Silva, D.N., Bortolao, V., Ramirez, M. and Carrico, J.A. (2019) Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.*, **47**, D188–D194.
- Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, J., Ostell, J., Pruitt, K.D. and Sayers, E.W. (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
- O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–745.
- Camargo, A.P., Call, L., Roux, S., Nayfach, S., Huntemann, M., Palaniappan, K., Ratner, A., Chu, K., Mukherjee, S., Reddy, T.B.K., et al. (2024) IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata. *Nucleic Acids Res.*, **52**, D164–D173.
- Lai, S., Jia, L., Subramanian, B., Pan, S., Zhang, J., Dong, Y., Chen, W.H. and Zhao, X.M. (2021) mMGE: a database for human metagenomic extrachromosomal mobile genetic elements. *Nucleic Acids Res.*, **49**, D783–D791.
- Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., et al. (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
- Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T. and Fujisawa, T. (2020) DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res.*, **48**, D45–D50.
- Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L. and Steinegger, M. (2022) Metagenome analysis using the Kraken software suite. *Nat. Protoc.*, **17**, 2815–2839.
- Cochrane, G., Bates, K., Apweiler, R., Tateno, Y., Mashima, J., Kosuge, T., Mizrahi, I.K., Schafer, S. and Fetchko, M. (2006) Evidence standards in experimental and inferential INSDC Third Party Annotation data. *OMICS*, **10**, 105–113.

31. Hauser, M., Steinegger, M. and Soding, J. (2016) MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, **32**, 1323–1330.
32. Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–143.
33. Robertson, J., Bessonov, K., Schonfeld, J. and Nash, J.H.E. (2020) Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb. Genom.*, **6**, 1099.
34. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
35. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
36. Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.
37. Russel, J., Pinilla-Redondo, R., Mayo-Munoz, D., Shah, S.A. and Sorensen, S.J. (2020) CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J.*, **3**, 462–469.
38. Nielsen, H., Teufel, F., Brunak, S. and von Heijne, G. (2024) SignalP: The Evolution of a Web Server. *Methods Mol. Biol.*, **2836**, 331–367.
39. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
40. Cantalapiedra, C.P., Hernandez-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
41. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
42. The Gene Ontology C. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
43. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
44. Schellenberger, J., Park, J.O., Conrad, T.M. and Palsson, B.O. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, **11**, 213.
45. Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D. and Koonin, E.V. (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.
46. Drula, E., Garron, M.L., Dogan, S., Lombard, V., Henrissat, B. and Terrapon, N. (2022) The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.*, **50**, D571–D577.
47. Buchfink, B., Reuter, K. and Drost, H.G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
48. Liu, B., Zheng, D., Zhou, S., Chen, L. and Yang, J. (2022) VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.*, **50**, D912–D917.
49. Alcock, B.P., Huynh, W., Chalil, R., Smith, K.W., Raphenya, A.R., Wlodarski, M.A., Edalatmand, A., Petkau, A., Syed, S.A., Tsang, K.K., *et al.* (2023) CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.*, **51**, D690–D699.
50. Blin, K., Shaw, S., Augustijn, H.E., Reitz, Z.L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B.R., Metcalf, W.W., Helfrich, E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.*, **51**, W46–W50.
51. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
52. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
53. Zieleszinski, A., Vinga, S., Almeida, J. and Karlowski, W.M. (2017) Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.*, **18**, 186.
54. Wickham, H. (2016) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY.
55. Wickham, H. (2011) The split-apply-combine strategy for data analysis. *J. Stat. Softw.*, **40**, 1–29.
56. Wang, X., Chen, L., Liu, W., Zhang, Y., Liu, D., Zhou, C., Shi, S., Dong, J., Lai, Z., Zhao, B., *et al.* (2023) TIMEDB: tumor immune micro-environment cell composition database with automatic analysis and interactive visualization. *Nucleic Acids Res.*, **51**, D1417–D1424.
57. Wang, R.H., Yang, S., Liu, Z., Zhang, Y., Wang, X., Xu, Z., Wang, J. and Li, S.C. (2024) PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic Acids Res.*, **52**, D756–D761.
58. Jia, W., Li, H., Li, S., Chen, L. and Li, S.C. (2020) Oviz-Bio: a web-based platform for interactive cancer genomics data visualization. *Nucleic Acids Res.*, **48**, W415–W426.
59. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
60. Terlouw, B.R., Blin, K., Navarro-Munoz, J.C., Avalon, N.E., Chevrete, M.G., Egbert, S., Lee, S., Meijer, D., Recchia, M.J.J., Reitz, Z.L., *et al.* (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.*, **51**, D603–D610.
61. Gorrie, C.L., Mirceta, M., Wick, R.R., Judd, L.M., Lam, M.M.C., Gomi, R., Abbott, I.J., Thomson, N.R., Strugnelli, R.A., Pratt, N.F., *et al.* (2022) Genomic dissection of *Klebsiella pneumoniae* infections in hospital patients reveals insights into an opportunistic pathogen. *Nat. Commun.*, **13**, 3017.