# IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata

Antonio Pedro Camargo[1,*], Lee Call[1], Simon Roux[1], Stephen Nayfach[1], Marcel Huntemann[1], Krishnaveni Palaniappan[1], Anna Ratner[1], Ken Chu[1], Supratim Mukherjeep[1], T. B. K. Reddy[1], I-Min A. Chen[1], Natalia N. Ivanova[1], Emiley A. Eloe-Fadrosh[1], Tanja Woyke[1], David A. Baltrus[2,3], Salvador Castañeda-Barba[4], Fernando de la Cruz[5], Barbara E. Funnell[6], James P. J. Hall[7], Aindrila Mukhopadhyay[8,9], Eduardo P. C. Rocha[10], Thibault Stalder[4], Eva Top[4], Nikos C. Kyrpides[1,*].

1. DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

2. School of Plant Sciences, University of Arizona, Tucson AZ, USA.

3. School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson AZ, USA.

4. Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA.

5. Instituto de Biomedicina y Biotecnología de Cantabria (Consejo Superior de Investigaciones Científicas – Universidad de Cantabria), Cantabria, Spain.

6. Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 1M1, Canada.

7. Department of Evolution, Ecology and Behaviour, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool L69 7ZB, UK.

8. Joint BioEnergy Institute, Emeryville, CA 94608, USA.

9. Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

10. Institut Pasteur, Université de Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, Paris, France.

* Correspondence should be addressed to: antoniop.camargo@lbl.gov, nckyrpides@lbl.gov.

## Abstract

Plasmids are mobile genetic elements found in many clades of Archaea and Bacteria. They drive horizontal gene transfer, impacting ecological and evolutionary processes within microbial communities, and hold substantial importance in human health and biotechnology. To support plasmid research and provide scientists with data of an unprecedented diversity of plasmid sequences, we introduce the IMG/PR database, a new resource encompassing 699,973 plasmid sequences derived from genomes, metagenomes, and metatranscriptomes. IMG/PR is the first database to provide data of plasmid that were systematically identified from diverse microbiome samples. IMG/PR plasmids are associated with rich metadata that includes geographical and ecosystem information, host taxonomy, similarity to other plasmids, functional annotation, presence of genes involved in conjugation and antibiotic resistance. The

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

database offers diverse methods for exploring its extensive plasmid collection, enabling users to navigate plasmids through metadata-centric queries, plasmid comparisons, and BLAST searches. The web interface for IMG/PR is accessible at https://img.jgi.doe.gov/pr. Plasmid metadata and sequences can be downloaded from https://genome.jgi.doe.gov/portal/IMG_PR.

## Introduction

Plasmids are ubiquitous mobile genetic elements that are present in several lineages of Archaea and Bacteria. These elements exhibit remarkable diversity, including varying length, mechanisms of mobility, strategies of replication, genetic makeup, copy number, and host range (1). Plasmids hold a central role in the evolution and ecology of microorganisms owing to their importance in driving horizontal gene transfer (HGT). Due to their genetic plasticity and capacity to move between different cells, plasmids allow cellular organisms to acquire genes from a communal genetic pool, rather than relying solely on vertical inheritance. This dynamic promotes the interchange of genetic material among distantly related lineages, facilitating adaptation to environmental pressures (2–5). Beyond their critical role in natural biological communities, plasmids also hold significant clinical value (6), illustrated by their involvement in disseminating antibiotic resistance genes and virulence factors (7, 8). Furthermore, plasmids find application in biotechnology, serving as indispensable tools for genetic manipulation.

To support plasmid research, scientists often leverage plasmid sequences from public databases. The primary sources of plasmid sequence data are INSDC repositories (9), such as NCBI's GenBank (10), which encompass the vast majority of Bacteria and Archaea genomes. However, these are general-purpose sequence databases and lack tools that allow targeted queries valuable for plasmid research, such as the presence of genes involved in conjugation or antibiotic resistance. Moreover, the determination of a sequence as a plasmid or not in these databases is based on the submitter's assessment, resulting in erroneous classification of chromosome sequences as plasmids. To tackle these issues, plasmid-specific databases such as pATLAS (11) PLSDB (12, 13), and COMPASS (14) provide data such as replicon and MOB typing, antibiotic resistance genes, and virulence genes. Furthermore, they offer users with more reliable sets of plasmid sequences, obtained by curating plasmids from GenBank or RefSeq to remove chromosomal sequences and redundant plasmids. However, because they source plasmids exclusively from RefSeq (15) or GenBank, these plasmid databases are mostly limited to single organism assemblies and do not include plasmid sequences from natural environments, present in metagenomic and metatranscriptomic assemblies. This constraint restricts the representation of plasmid diversity to a relatively small fraction of microbial diversity, excluding the majority of uncultivated microorganisms.

In recent years, the automated identification of viral genomes within metagenomic and metatranscriptomic data has significantly altered our understanding of virosphere diversity, enabling the detection of numerous major lineages that would have otherwise eluded detection (16–20). Databases such as IMG/VR (21) leverage this data, enabling researchers to explore uncultivated viruses identified across thousands of samples, thereby supporting virus-related research. In contrast, automatic large-scale detection of plasmids from sequencing data of natural environments is still limited to a few studies (22–26), and no database currently provides automatically identified plasmids from microbiomes.

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

To address this issue, we have developed IMG/PR, a comprehensive database of plasmid sequences derived from the Integrated Microbial Genomes & Microbiomes (IMG/M) system (27). IMG/PR amasses the most extensive collection of publicly available plasmids, systematically identified across several thousand genomes, metagenomes, and metatranscriptomes using the geNomad tool (28). These plasmids are presented within a structured framework encompassing functional annotation and extensive metadata, which includes host taxonomy, completeness, similarity to other plasmids, geographic coordinates, and ecosystem information, allowing users to investigate plasmid function and ecology thoroughly. Users can explore IMG/PR data by searching through the metadata associated with each plasmid or by performing sequence-based BLAST searches, making it easy to identify relevant plasmids. We anticipate that IMG/PR will become a primary resource for plasmid research, allowing researchers to systematically assess environmental plasmids on a large scale for the first time.

## Methods

### Identification of plasmid sequences in IMG/M datasets

Plasmid sequences were identified through screening of all assembled datasets fetched from the IMG/M database as of 2022-04-10. This collection included 28,865 metagenomes, 7,258 metatranscriptomes, 83,858 isolate genomes of Bacteria and Archaea, 4,342 single amplified genomes (SAGs), and 10,499 metagenome-assembled genomes (MAGs). A minimum contig length requirement of 2 kb (for contigs with direct or inverted terminal repeats) or 4 kb (for the remaining contigs) was applied for metagenomes, isolate genomes, SAGs, and MAGs. For metatranscriptomes, a length cutoff of 2 kb was enforced throughout. Identification of plasmid sequences was carried out using geNomad (version 1.1.0), and a composition-based score calibration was used to set the false discovery rate to 2% (parameters: '--enable-score-calibration --max-fdr 0.02'). The identified putative plasmids underwent additional filtration, applying criteria that provide orthogonal support for the initial classification performed with geNomad. This step aimed to exclude sequence fragments for which the plasmid origin remained ambiguous, potentially representing segments of genomic islands. To meet inclusion criteria in IMG/PR, sequences were required to score at least five points based on the following criteria, each manually weighted according to their prevalence in sequences exhibiting high similarity to reference plasmids:

- Share at least 50% of their genes with a reference plasmid: **1 point**.
- Match at least two CRISPR spacers from a database of spacers identified in the genomes of Bacteria and Archaea (see the "*Assignment to host taxa*" section): **2 points**.
- Exhibit direct terminal repeats (DTRs): **2 points**.
- Encoding at least one plasmid hallmark gene (excluding relaxases), as defined by geNomad: **2 points**.
- Encode a relaxase or an origin of transfer (see the "*Annotation of plasmid sequences*" section): **3 points**.

To identify genes shared with reference plasmids, gene prediction was performed with prodigal-gv (version 2.10.0, available at https://github.com/apcamargo/prodigal-gv) (29) for both IMG and reference sequences and DIAMOND (version 2.1.1) (30) was then used to perform protein alignment. Then, for each

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

pair of sequences, shared genes were identified as reciprocal best hits with alignment coverage ≥50% for both the query and target proteins (code available at https://github.com/apcamargo/bioinformatics-snakemake-pipelines/tree/main/contig-aai-pipeline). The reference plasmids used for this comparison were obtained from PLSDB (v. 2021_06_23_v2) for bacterial plasmids and RefSeq (retrieved on 2022-06-16, query: 'archaea[filter] AND refseq[filter] AND plasmid[filter]') for archaeal plasmids. We identified DTRs by finding sequences with exact matches of at least 21 bp at both ends. To prevent DTRs from being called in repetitive regions, we detected low complexity sequence segments using dustmasker (version 1.0.0, parameters: '-level 40'). DTRs were disregarded if the low complexity region constituted 50% or more of their length.

Putative genomic islands were excluded from the dataset by identifying sequences that aligned to chromosomes of Bacteria and Archaea across ≥90% of their length. Chromosome sequences were sourced from RefSeq (on 2023-04-24) by retrieving all nucleotide sequences from bacterial and archaeal genomes with at least 500 kb and without the word "plasmid" in their headers. Alignments were performed using blastn (parameters: '-task megablast -evalue 1e-5 -max_target_seqs 20000') (31).

### Annotation of plasmid sequences

In addition to the functional information provided by the IMG/M annotation pipeline (32), plasmids in IMG/PR were further annotated via the identification of protein components of conjugation systems, origins of transfer, and antibiotic resistance genes. Proteins that are part of conjugation systems (relaxases, *VirB4* T4SS ATPases, T4CPs, and other T4SS components) were detected by using hmmsearch (version 3.3.2, parameters: '-E 1e-3') (33) to search the CONJscan (34) HMM models against the sequences of the proteins encoded by IMG/PR plasmids, with a minimum HMM coverage of 50% required. The determination of complete conjugative systems was based on the presence of the minimum set of components outlined in the CONJscan model definitions. Origins of transfer (*oriT*) were identified by searching the IMG/PR plasmid sequences against a database of 91 previously described *oriT* sequences (35) using blastn (parameters: '-task blastn-short -outfmt '6 std qlen slen qseq sseq' -num_threads 64 -dust no'), with a minimum target coverage of 20% required. Antibiotic resistance genes (ARGs) were annotated using hmmsearch (parameters: '--cut_ga') to search Resfams (36) HMMs against IMG/PR proteins.

### Assignment to host taxa

The host taxa of plasmids identified within isolate assemblies or SAGs were determined through guilt-by-association, whereby these plasmids were assigned to the taxon of the genome in which they were identified. Plasmids originating from MAGs, metagenomes, and metatranscriptomes were assigned to host taxa by indirect association, leveraging matches to a database of 4.8 million unique CRISPR spacers (21) (available at https://portal.nersc.gov/cfs/m342/crisprDB), obtained from a collection of 1.6 million bacterial and archaeal genomes retrieved from NCBI GenBank (release 242) and multiple MAG datasets (37–40). Plasmids were aligned to the spacer database using blastn (version 2.13.0+, parameters: '-max_target_seqs=1000 -word_size=8 -dust=no') (41), with only alignments covering at least 25 base pairs, allowing for a maximum of 1 mismatch, and spanning ≥95% of the spacer length being considered. Plasmids that matched at least two CRISPR spacers were then assigned to a host taxa using a majority vote

rule, where the most specific taxon that represented >70% of the CRISPR spacer hits was selected as the host. To ensure uniformity throughout the entire database, all isolate genome assemblies and SAGs from IMG/M, as well as genomes within the CRISPR spacer database, were assigned to taxonomic lineages as defined in the GTDB database (release 207) (42–44) using GTDB-Tk (version 2.1.0) (45).

### Clustering of plasmids into PTUs

Plasmid sequences were clustered into plasmid taxonomic units (PTUs), which group sequences sharing a common genomic backbone with elevated average nucleotide identity (ANI) (46–48). Due to sequence fragmentation, which precludes the identification of complete backbone sequences, PTUs were delimited in three steps: (1) clustering of complete plasmids, (2) recruitment of fragments to clusters of complete plasmids, (3) clustering of the remaining fragments. Putatively complete plasmids were identified either by the presence of DTRs, or by alignment to complete reference plasmids (see the "*Identification of plasmid sequences in IMG/M datasets*" section) across the entirety of the length of both the query sequence and the reference with high ANI (query and reference coverage ≥ 99%, ANI ≥ 95%). These complete plasmids were used to construct a graph connecting sequences meeting the aligned fraction (AF) and ANI criteria (AF of the shorter plasmid in the pair ≥ 50%, ANI ≥ 70%) (46). The graph was then subjected to clustering using the Leiden algorithm (49). In the second clustering step, plasmid fragments were recruited to the initial clusters if they aligned with a complete plasmid across ≥85% of their length ANI ≥ 70%. Each fragment was assigned to a single cluster. The remaining fragments were used to build a new graph with stricter connection criteria (AF of the shorter sequence ≥ 85%, ANI ≥ 70%), which underwent clustering using the Leiden algorithm. Throughout the process, blastn (parameters: '-task megablast -evalue 1e-5 -max_target_seqs 20000') was used for alignments, ANI and AF were calculated by aggregating high-scoring segment pairs between pairs of plasmids (code available at https://bitbucket.org/berkeleylab/checkv/src/master/scripts/anicalc.py), and pyLeiden (available at https://github.com/apcamargo/pyleiden, parameters: '-n 5 -r 1.2') was employed for clustering. During clustering, graph edges were weighted by pairwise ANI and AF ($weight = AF \times ANI$).

## Results

IMG/PR is a public database for the exploration and analysis of plasmid sequences identified in datasets integrated into the IMG/M system. IMG/PR provides additional metadata in addition to those offered by IMG/M, allowing users to investigate plasmid sequences in regards to their geographical distribution, ecosystem attributes, host taxonomy, and functions such as conjugation and antibiotic resistance genes.

### IMG/PR database composition

IMG/PR contains a total of 699,973 plasmid sequences, which were automatically identified across thousands of metagenomes, isolate genomes, MAGs, SAGs, and metatranscriptomes (Figure 1A). Among those, 154,680 (22.1%) were determined to be complete, either by the presence of direct terminal repeats or by complete bidirectional alignment with complete reference plasmids (Figure 1A). Notably, complete plasmids identified within metagenomes tend to be shorter than those of isolate genomes (median lengths of 5.2 kb and 18.7 kb, respectively; Figure 1B). This difference can be attributed to the challenges associated with assembling long contigs from short-read metagenomic data, and possible selection biases

in cultivated microorganisms. Comparison with reference sequences was also used to determine that 287,783 (41.1%) of the IMG/PR sequences display high similarity to references (defined as having at least 50% of their length aligned with a reference plasmid), underscoring the large amount of novel plasmids in the database (Figure 1A). Among the plasmids that do not exhibit high similarity to any reference, 77,756 (18.9%) are considered complete, representing a notable expansion of the existing catalogue of complete plasmids.

The plasmids within IMG/PR are organised in 214,950 PTUs, which represent clusters of interconnected plasmids sharing genetic backbones. Most PTUs contain a single sequence (137,282, or 63.9%) and only 8,844 PTUs (4.1%) comprise ten plasmids or more (Figure 1C). As plasmids often evolve by the gain and loss of accessory genes while maintaining a conserved genetic backbone, the clustering of plasmids into PTUs organises plasmid diversity and allows researchers to assess, for example, the evolution, host range, and environmental distribution of specific groups of plasmids. By evaluating in which ecosystems the plasmids within PTUs with at least ten members were detected, we found that, although most of these PTUs (66.6%) are specific to a single ecosystem, a substantial fraction contain members found in multiple ecosystems (Figure 1D). Among those, the most common combinations encompassed PTUs present in human-associated and animal-associated samples, as well as PTUs found in human-associated samples and metagenomes from engineered environments, such as wastewater.

### IMG/PR includes plasmids from diverse ecosystems and geographic regions

One distinctive feature of IMG/PR is its incorporation of not only plasmids from isolates but also plasmid sequences automatically identified within metagenomes and metatranscriptomes. This inclusion markedly augments the sequence diversity within IMG/PR (Figure 2A, left), emphasising the potential of plasmids existing in natural ecosystems for the advancement of plasmid research.

IMG/PR sources metadata from the Genomes OnLine Database (GOLD) (50) to provide geographical and ecosystem context for plasmids identified within metagenomes and metatranscriptomes. Geographic coordinate information reveals that IMG/PR spans plasmids identified worldwide, with the majority originating from samples in North America and Europe (Figure 2B). In terms of the ecosystems where these environmental plasmids were identified, sequences from human-associated communities, such as the human gut microbiome, outnumber those from other ecosystems. However, the number of PTUs detected in human-associated samples is lower compared to other ecosystems (Figure 2A, centre), such as soil, and displays a lower growth tendency (Figure 2A, right). This underscores the value of IMG/PR in providing a comprehensive resource of plasmids from less explored ecosystems.

### IMG/PR plasmids are assigned to varied host taxa

A set of 279,412 IMG/PR plasmids (39.9% of the total) were assigned to host taxa within Archaea (4 phyla; 7 classes) and Bacteria (45 phyla; 94 classes), providing additional biological context for these sequences (Figures 1A, 3A). Host assignment was performed using two distinct approaches (as detailed in "*Methods*"): plasmids identified in genomes of isolates or in SAGs were assigned using a guilt-by-association strategy to the taxa of the corresponding genome (49.5% of the assigned plasmids), while the remaining plasmids linked to predicted host taxa based on matches to CRISPR spacer sequences (50.5% of the assigned plasmids).

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

The overwhelming majority of the plasmids with host information were assigned to Bacteria (99.6%), particularly to the *Proteobacteria*, *Firmicutes*, *Bacteroidota*, and *Actinobacteriota* phyla. The large number of *Bacteroidota* plasmids (n=41,318, grouped into 4,004 PTUs) in IMG/PR is noteworthy, as this phylum, highly prevalent in the human gut, is relatively scarce in other databases (319 sequences in PLSDB v. 2021_06_23_v2) (46). Analysis of the proportion of plasmid sequences showing similarity to reference sequences reveals considerable variation across distinct host taxa (Figure 3B) and highlights that even prevalent taxa encompass substantial novelty.

IMG/PR facilitates the integration of host information with diverse metadata for comprehensive plasmid assessment. This enables, for instance, the investigation of the ecosystem distribution of major taxa (Figure 3C), which reveals that while plasmids of *Gammaproteobacteria* are found across various environments, plasmids assigned to taxa such as *Bacteroidota* and *Cyanobacteria* are primarily limited to specific niches (human/animal-associated and freshwater environments, respectively).

### IMG/PR allows investigation of plasmid function, conjugative potential, and antibiotic resistance

All plasmid sequences in IMG/PR are linked to an IMG/M scaffold and have been annotated through the IMG Annotation Pipeline. As a result, IMG/PR provides the genomic coordinates and sequences of all genes encoded by the plasmids, along with associated Pfam (51), COG (52), TIGRFAM (53), and KEGG (54) Orthology accessions (Figure 4A). This enables users to readily interrogate plasmid functions using established classification systems.

Beyond standard annotation, IMG/PR plasmids underwent additional analysis using domain-specific databases. As a result, IMG/PR incorporates information regarding the presence of elements constituting the conjugation machinery and of antibiotic resistance genes, both of which hold significant importance in plasmid research.

Across the database, hundreds of thousands of plasmids harbour components of the conjugation system (Figure 4B), including the relaxase (MOB), *VirB4* type IV secretion system (T4SS) ATPase, type 4 coupling protein (T4CP), other T4SS proteins, and the origin of transfer (*oriT*). Conjugative T4SSs enable mating pair formation (MPF) and, by extension, plasmid transfer via conjugation (55). Thus, plasmids encoding the complete set of genes required for MPF formation, a relaxase, and a T4CP are inferred to be capable of autonomous conjugation. IMG/PR encompasses 57,247 such plasmids (grouped into 19,994 PTUs), spanning eight distinct MPF types with unique taxonomic compositions (Figure 4C).

In the context of antibiotic resistance, IMG/PR includes 33,815 plasmids (9,440 PTUs) encoding at least one predicted antibiotic resistance gene (ARG, Figure 4D). Among these, 20,033 plasmids (59.2%) encode a relaxase, and 4,851 (14.3%) possess all genes essential for autonomous conjugation, underscoring IMG/PR's utility in assessing potential antibiotic resistance gene dissemination via conjugation. To further analyse the distribution of ARGs, IMG/PR enables cross-referencing of ARGs, host taxonomy, and ecosystem data to evaluate the mechanisms of plasmid-encoded antibiotic resistance across diverse taxa and environments (Figure 4E).

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

### Functionalities for exploration and analysis of plasmids in IMG/PR

As illustrated in the preceding sections, IMG/PR offers an extensive collection of plasmids within a comprehensive framework that supplies abundant metadata and tools that enable exploration of plasmid diversity. All plasmids in IMG/PR are systematically organized in PTUs, that group related plasmids, and are linked to an IMG/M scaffold from an isolate genome, SAG, MAG, metagenome, or metatranscriptome. In addition, each plasmid has a series of attributes (Figure 5, left), derived either from the IMG/M system — encompassing sequence length, gene count, Pfam domains, ecosystem, and geographical location — or specific to IMG/PR — including plasmid completeness, sequence topology (whether the sequence contains DTRs, ITRs, or concatemers), similarity to a reference plasmid, geNomad score (indicating the confidence level of geNomad's plasmid prediction), host taxonomy, presence of conjugation genes, presence of an origin of transfer, and presence of ARGs.

To facilitate user interaction with the data for addressing their research questions, IMG/PR offers multiple tools that enable distinct approaches to data interrogation (Figure 5, right). First, users can search the database by querying plasmids based on their attributes, allowing the identification of plasmids that match user-defined criteria. This attribute-centric search approach allows cross-referencing of distinct attribute types, making it possible to perform complex searches to answer their biological questions. Second, users can compare plasmids within IMG/PR using the 'find similar plasmids' tool, which leverages geNomad markers to compare plasmids based on their gene content, making it possible to identify even distantly related plasmids that share selected genes. Lastly, the BLAST functionality enables users to search IMG/PR using their own protein and nucleotide sequences, in order to identify plasmids that exhibit similarity to theirs within the database.

## Conclusions

The investigation of plasmids is pivotal for understanding the evolution, ecology, and molecular biology of microorganisms. Nevertheless, the vast majority of plasmid sequences found in databases come from a small a fraction of microbial diversity, hindering comprehensive explorations of plasmid diversity and function. To address this issue, we present IMG/PR, a novel database of plasmid sequences that are sourced not only from genomes of Bacteria and Archaea but also from metagenomes and metatranscriptomes of natural environments, thus encompassing a broader spectrum of plasmid diversity. IMG/PR provides plasmid sequences alongside functional, taxonomic, and ecological metadata, as well as tools that empower users to discover plasmids based on their attributes or sequences. We anticipate that IMG/PR will serve as a crucial resource for plasmid research, enabling scientists to explore plasmids from previously uncultivated microorganisms across a wide and diverse range of natural ecosystems. Specifically, we believe that IMG/PR will prove invaluable in supporting research on the dynamics of HGT through conjugation under diverse conditions, identifying potential hotspots for the dissemination of antibiotic resistance genes in epidemiological studies, and aiding in the development of novel treatments targeting antibiotic resistance genes and virulence factors. As a result, we believe that IMG/PR will have far-reaching implications in multiple scenarios, including scientific research, epidemiology, and healthcare.

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

## Data availability

The IMG/PR web interface is accessible through: https://img.jgi.doe.gov/pr/. Metadata, nucleotide, and protein sequences associated with IMG/PR plasmids can be downloaded from: https://genome.jgi.doe.gov/portal/IMG_PR.

## Funding

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), and the National Energy Research Scientific Computing Center (NERSC) (https://ror.org/05v3mvq14), is supported by the DOE Office of Science User Facilities operated under Contract No. DE-AC02-05CH11231. Computational resources were partly provided by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. JPJH is supported by an MRC Career Development Award (MR/W02666X/1).

## References

1. Rodríguez-Beltrán,J., DelaFuente,J., León-Sampedro,R., MacLean,R.C. and San Millán,Á. (2021) Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.*, **19**, 347–359.

2. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

3. de la Cruz,F. and Davies,J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, **8**, 128–133.

4. Koonin,E.V. (2016) Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, **5**, 1805.

5. Haudiquet,M., De Sousa,J.M., Touchon,M. and Rocha,E.P.C. (2022) Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos. Trans. R. Soc. B Biol. Sci.*, **377**, 20210234.

6. San Millan,A. (2018) Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends Microbiol.*, **26**, 978–985.

7. Sugino,Y. and Hirota,Y. (1962) Conjugal fertility associated with resistance factor R in *Escherichia coli*. *J. Bacteriol.*, **84**, 902–910.

8. Nassif,X., Fournier,J.M., Arondel,J. and Sansonetti,P.J. (1989) Mucoid phenotype of Klebsiella pneumoniae is a plasmid-encoded virulence factor. *Infect. Immun.*, **57**, 546–552.

9. Karsch-Mizrachi,I., Takagi,T., Cochrane,G., and on behalf of the International Nucleotide Sequence Database Collaboration (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **46**, D48–D51.

10. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2017) GenBank. *Nucleic Acids Res.*, **45**, D37–D42.

11. Jesus,T.F., Ribeiro-Gonçalves,B., Silva,D.N., Bortolaia,V., Ramirez,M. and Carriço,J.A. (2019) Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.*, **47**, D188–D194.

12. Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2019) PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, **47**, D195–D202.

13. Schmartz,G.P., Hartung,A., Hirsch,P., Kern,F., Fehlmann,T., Müller,R. and Keller,A. (2022) PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.*, **50**, D273–D278.

14. 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.

15. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,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–D745.

16. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G.A., Thomas,A.D., Huntemann,M., Mikhailova,N., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering Earth's virome. *Nature*, **536**, 425–430.

17. Roux,S., Krupovic,M., Daly,R.A., Borges,A.L., Nayfach,S., Schulz,F., Sharrar,A., Matheus Carnevali,P.B., Cheng,J.-F., Ivanova,N.N., *et al.* (2019) Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.*, **4**, 1895–1906.

18. Schulz,F., Roux,S., Paez-Espino,D., Jungbluth,S., Walsh,D.A., Denef,V.J., McMahon,K.D., Konstantinidis,K.T., Eloe-Fadrosh,E.A., Kyrpides,N.C., *et al.* (2020) Giant virus diversity and host interactions through global metagenomics. *Nature*, **578**, 432–436.

19. Edgar,R.C., Taylor,J., Lin,V., Altman,T., Barbera,P., Meleshko,D., Lohr,D., Novakovsky,G., Buchfink,B., Al-Shayeb,B., *et al.* (2022) Petabase-scale sequence alignment catalyses viral discovery. *Nature*, **602**, 142–147.

20. Neri,U., Wolf,Y.I., Roux,S., Camargo,A.P., Lee,B., Kazlauskas,D., Chen,I.M., Ivanova,N., Zeigler Allen,L., Paez-Espino,D., *et al.* (2022) Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*, **185**, 4023-4037.e18.

21. Camargo,A.P., Nayfach,S., Chen,I.-M.A., Palaniappan,K., Ratner,A., Chu,K., Ritter,S.J., Reddy,T.B.K., Mukherjee,S., Schulz,F., *et al.* (2023) IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.*, **51**, D733–D743.

22. Jørgensen,T.S., Xu,Z., Hansen,M.A., Sørensen,S.J. and Hansen,L.H. (2014) Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. *PLoS ONE*, **9**, e87924.

23. Antipov,D., Raiko,M., Lapidus,A. and Pevzner,P.A. (2019) Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.*, **29**, 961–968.

24. Yu,M.K., Fogarty,E.C. and Eren,A.M. (2020) The genetic and ecological landscape of plasmids in the human gut. *bioRxiv*, 10.1101/2020.11.01.361691.

25. Stockdale,S.R., Harrington,R.S., Shkoporov,A.N., Khokhlova,E.V., Daly,K.M., McDonnell,S.A., O'Reagan,O., Nolan,J.A., Sheehan,D., Lavelle,A., *et al.* (2022) Metagenomic assembled plasmids of the human microbiome vary across disease cohorts. *Sci. Rep.*, **12**, 9212.

26. Conteville,L.C. and Vicente,A.C.P. (2022) A plasmid network from the gut microbiome of semi-isolated human groups reveals unique and shared metabolic and virulence traits. *Sci. Rep.*, **12**, 12102.

27. Chen,I.-M.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S.J., Webb,C., Wu,D., *et al.* (2023) The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.*, **51**, D723–D732.

28. Camargo,A.P., Roux,S., Schulz,F., Babinski,M., Xu,Y., Hu,B., Chain,P.S.G., Nayfach,S. and Kyrpides,N.C. (2023) Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*, 10.1038/s41587-023-01953-y.

29. 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.

30. Buchfink,B., Reuter,K. and Drost,H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.

31. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.*, **7**, 203–214.

32. Clum,A., Huntemann,M., Bushnell,B., Foster,B., Foster,B., Roux,S., Hajek,P.P., Varghese,N., Mukherjee,S., Reddy,T.B.K., *et al.* (2021) DOE JGI Metagenome Workflow. *mSystems*, **6**, e00804-20.

33. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Comput. Biol.*, **7**, e1002195.

34. Cury,J., Abby,S.S., Doppelt-Azeroual,O., Néron,B. and Rocha,E.P.C. (2020) Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. In de la Cruz,F. (ed), *Horizontal Gene Transfer: Methods and Protocols*, Methods in Molecular Biology. Springer US, New York, NY, pp. 265–283.

35. Ares-Arroyo,M., Coluzzi,C. and Rocha,E.P.C. (2023) Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Res.*, **51**, 3001–3016.

36. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.

37. Pasolli,E., Asnicar,F., Manara,S., Zolfo,M., Karcher,N., Armanini,F., Beghini,F., Manghi,P., Tett,A., Ghensi,P., *et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, **176**, 649-662.e20.

38. Nayfach,S., Roux,S., Seshadri,R., Udwary,D., Varghese,N., Schulz,F., Wu,D., Paez-Espino,D., Chen,I.-M., Huntemann,M., *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.

39. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P., *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.

40. Carter,M.M., Olm,M.R., Merrill,B.D., Dahan,D., Tripathi,S., Spencer,S.P., Yu,F.B., Jain,S., Neff,N., Jha,A.R., *et al.* (2023) Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell*, **186**, 3111-3124.e13.

41. 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.

42. Parks,D.H., Chuvochina,M., Waite,D.W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.

43. Parks,D.H., Chuvochina,M., Chaumeil,P.-A., Rinke,C., Mussig,A.J. and Hugenholtz,P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.

44. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.

45. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, **38**, 5315–5316.

46. Redondo-Salvo,S., Fernández-López,R., Ruiz,R., Vielva,L., De Toro,M., Rocha,E.P.C., Garcillán-Barcia,M.P. and De La Cruz,F. (2020) Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.*, **11**, 3602.

47. Redondo-Salvo,S., Bartomeus-Peñalver,R., Vielva,L., Tagg,K.A., Webb,H.E., Fernández-López,R. and De La Cruz,F. (2021) COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics*, **22**, 390.

48. Garcillán-Barcia,M.P., Redondo-Salvo,S. and De La Cruz,F. (2023) Plasmid classifications. *Plasmid*, **126**, 102684.

49. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.

50. Mukherjee,S., Stamatis,D., Li,C.T., Ovchinnikova,G., Bertsch,J., Sundaramurthi,J.C., Kandimalla,M., Nicolopoulos,P.A., Favognano,A., Chen,I.-M.A., *et al.* (2023) Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.*, **51**, D957–D963.

51. 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.

*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*

52. 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.

53. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.

54. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

55. Cabezón,E., Ripoll-Rozada,J., Peña,A., De La Cruz,F. and Arechaga,I. (2014) Towards an integrated model of bacterial conjugation. *FEMS Microbiol. Rev.*, 10.1111/1574-6976.12085.

## Table and Figure Legends

**Figure 1. (A)** IMG/PR composition in terms of data source (isolate genome, metagenome, metatranscriptome, MAG, or SAG), plasmid completeness, similarity to reference plasmids, and availability of host data. Each square represents 1% of the plasmids in IMG/PR. MAG: metagenome-assembled genome; SAG: single-amplified genome; DTRs: direct terminal repeats; ITRs: inverted terminal repeats. **(B)** Distribution of plasmid lengths (in kilobases) for putatively complete plasmids detected in isolate genomes and metagenomes. The vertical axis is presented in logarithmic scale. **(C)** Distribution of PTU sizes (number of plasmids within a PTU) in the IMG/PR PTUs with at least 10 members (n=8,844). Individual data points are shown on the right of the boxplot. The vertical axis is presented in logarithmic scale. **(D)** UpSet plot illustrating the distribution of ecosystem occurrences for PTUs containing members identified in metagenomes or metatranscriptomes. Only PTUs with a minimum of 10 members found in metagenomes or metatranscriptomes with geographic coordinate information are displayed (n=6,490). To determine the presence of a PTU within a specific ecosystem, at least two associated plasmids assigned to that PTU needed to be detected in the respective ecosystem. Box plots show the median values (middle line), interquartile range (box boundaries), and 1.5 times the interquartile range (whiskers).

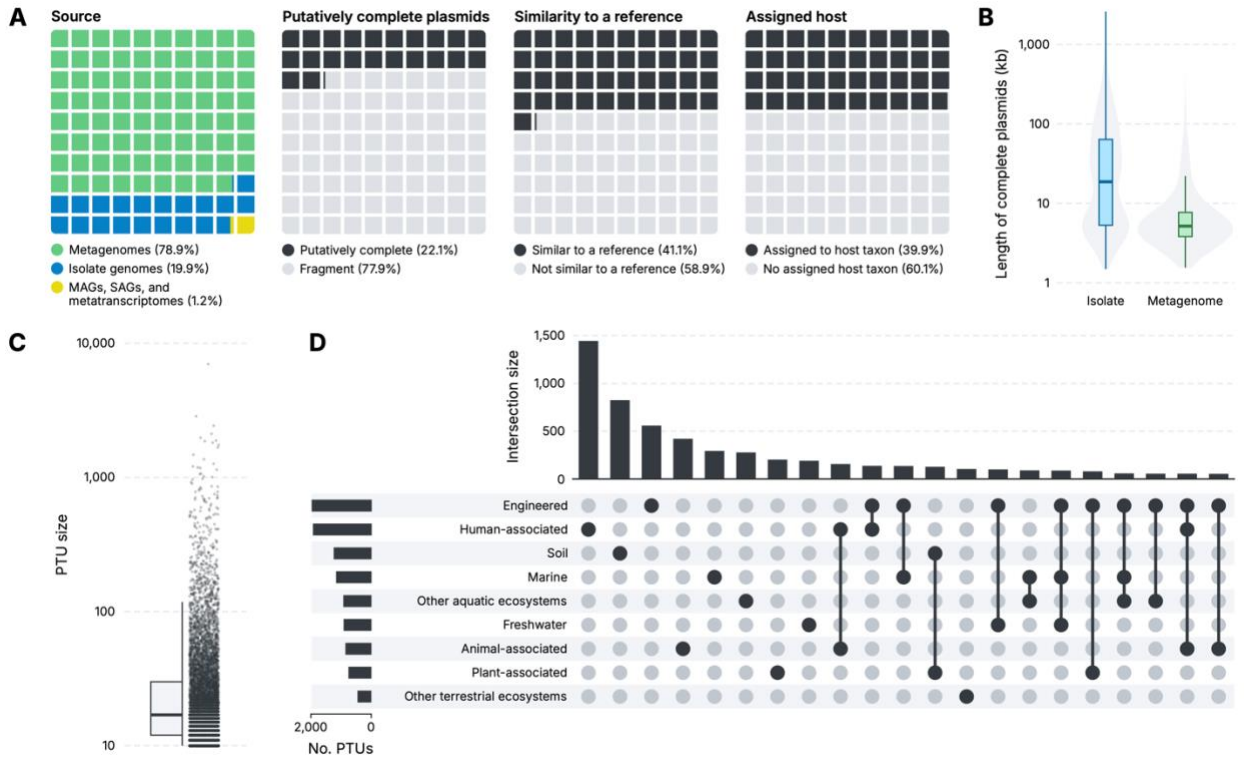
**Figure 2. (A)** PTU rarefaction curves and their corresponding growth tendencies for specific subsets of IMG/PR. The left panel shows the rarefaction of PTUs as a function of the number of plasmids detected in isolates (blue) and metagenomes (green). The centre panel shows PTU rarefaction curves for different types of ecosystems, considering only plasmids identified in metagenomes or metatranscriptomes. The table in the right panel displays the growth tendencies (gamma) of the rarefaction curves from the left and middle panels. The gamma values were derived by fitting the Heaps' law function to each curve and indicate whether a given curve is under rapid growth (gamma close to 1) or nearing saturation (gamma close to 0). Each rarefaction curve was constructed by averaging 100 sampling processes performed on random permutations of the data. **(B)** Geographical distribution of IMG/PR sequences at the PTU level (n=104,065) based on IMG/M metagenomes and metatranscriptomes with geographic coordinate information (n=11,644). The area of the circles on the map is proportional to the number of PTUs in each specific region, while their colours represent the number of metagenome and metatranscriptome samples in which the sequences were detected (in logarithmic scale). Samples were grouped within defined intervals along the longitude and latitude. The map is presented using the Mollweide projection.

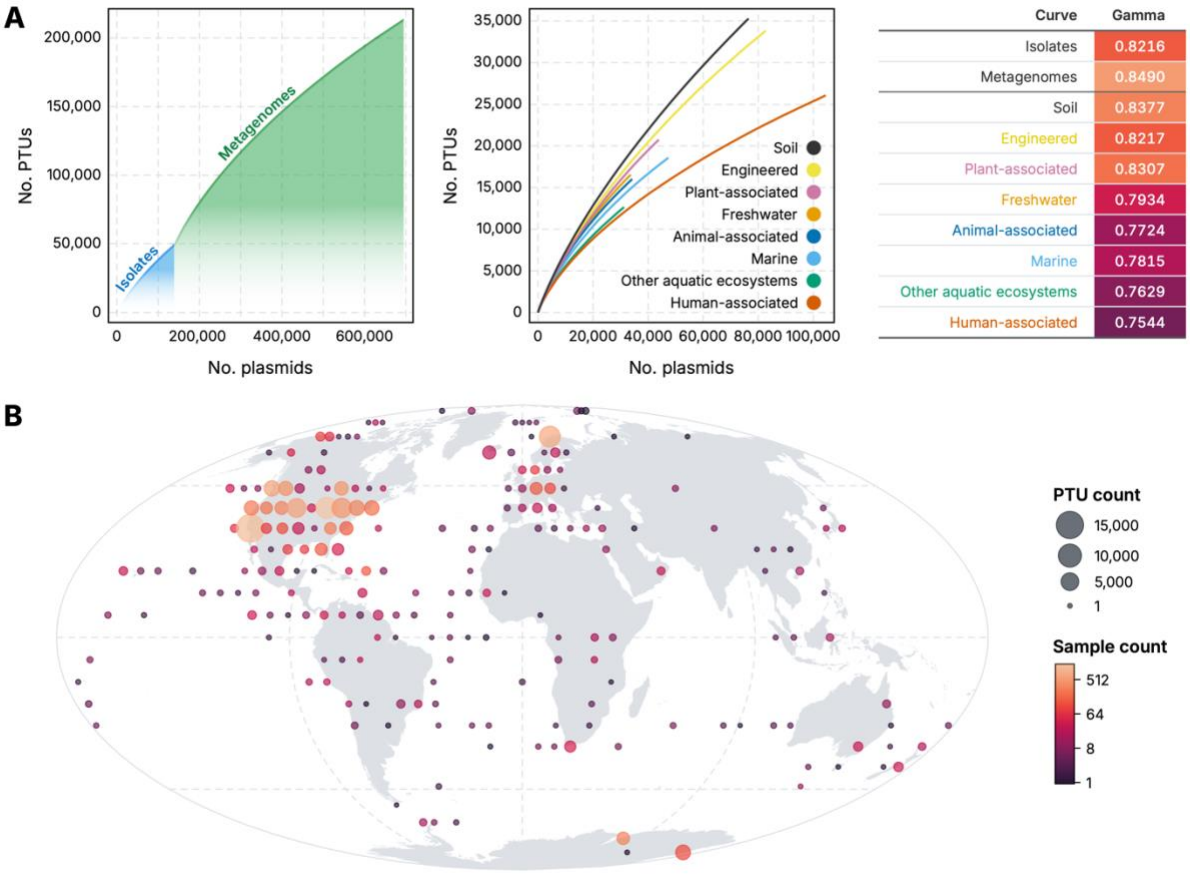
**Figure 3. (A)** Plasmid sequences assigned to bacterial (top cladogram) and archaeal (bottom cladogram) putative hosts. Leaves in the cladograms on the left show bacterial and archaeal classes. The bars in the centre represent the number of plasmids assigned to each class (shown in logarithmic scale) and the bars on the right show the relative frequencies for each of the host assignment methods within each class. Only classes with a minimum of five assigned plasmid sequences are shown in the tree. The remaining classes are grouped in "Other Bacteria" and "Other Archaea". The topology of the cladograms were derived from phylogenetic trees retrieved from GTDB (release 207). **(B)** Percentage of plasmid sequences exhibiting high similarity to reference plasmids across different host taxa. The top five most prevalent host classes are presented individually, while the remaining are grouped under the "Other" category. **(C)** Taxonomic composition of the hosts of plasmids identified in metagenomes and metatranscriptomes of the six ecosystem categories with the most plasmids. Only the five most common classes within each ecosystem category are shown, the remaining are grouped under "Other". The area of the pie charts is proportional to the number of plasmid sequences in the ecosystem.

**Figure 4. (A)** Most prevalent functional annotations found in IMG/PR plasmids. The length of each bar corresponds to the number of sequences encoding at least one gene annotated with the respective accession. The five most common Pfam, TIGRFAM, COG, and KEGG Ortholog accessions are presented. **(B)** Number of distinct plasmids (dark bars) and PTUs (light bars) encoding different components of the conjugation machinery. MOB: relaxase; T4SS: type IV secretion system; T4CP: type IV coupling protein; oriT: origin of transfer. **(C)** Taxonomic profile of sequences encoding various types of MPFs (mating pair formation) systems. Only sequences encoding all proteins required for autonomous conjugation, as defined by CONJscan, and taxonomically assigned to host taxa at the class level, are depicted. Host classes with fewer than 50 sequences encoding a particular MPF type are grouped under the "Other" category. The numbers above the bars indicate the total count of plasmid sequences within each MPF type. **(D)** Number of distinct plasmids (dark bars) and PTUs (light bars) encoding antibiotic resistance genes. Genes were grouped based on their resistance mechanisms, as classified in the Resfams database. **(E)** Alluvial plot depicting the relationship between host phyla (rectangles on the left), antibiotic resistance mechanism (rectangles in the middle), and ecosystem (rectangles on the right) of antibiotic resistance genes encoded by plasmids identified in metagenomes or metatranscriptomes. The height of each rectangle reflects the number of antibiotic resistance in the respective category. Relationships between host phyla, antibiotic resistance mechanisms, and ecosystems are indicated by curved lines, their width corresponding to the number of antibiotic resistance genes. Curves are colour-coded based on the antibiotic resistance mechanism.

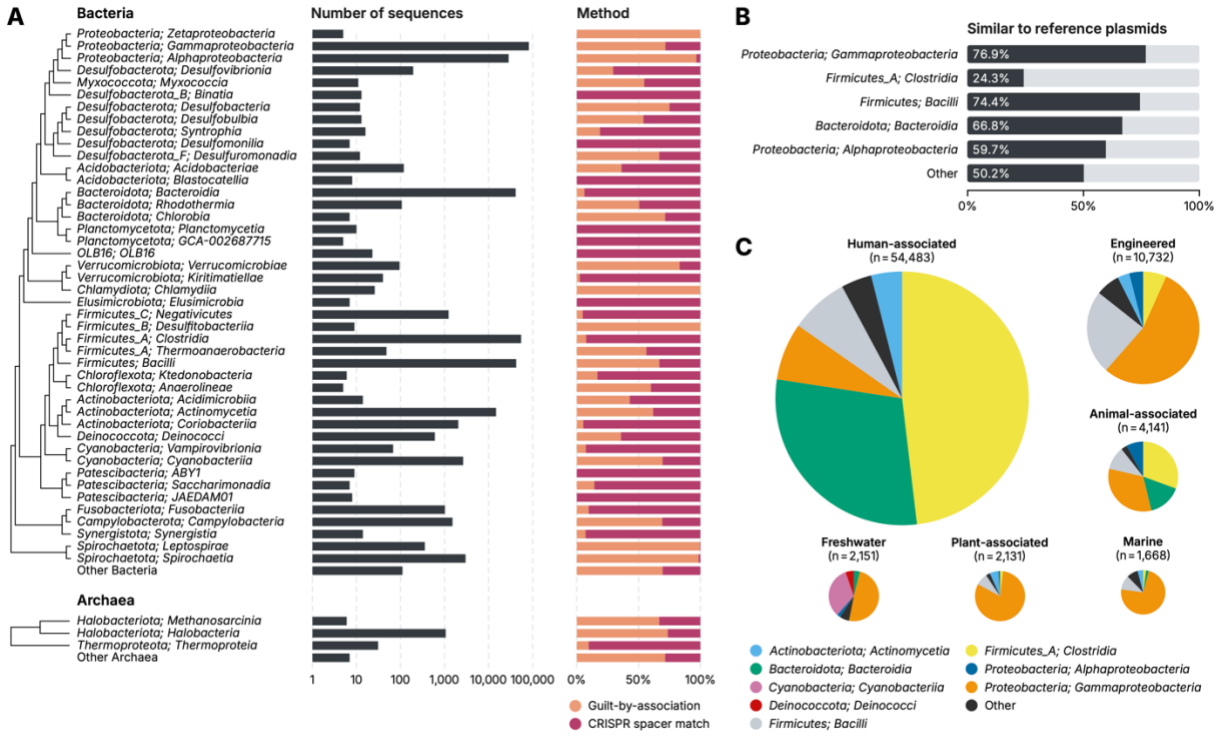
**Figure 5.** The plasmids in IMG/PR are organised into PTUs, which are clusters of related plasmid sequences. Each plasmid is linked to an IMG/M scaffold, its nucleotide sequence, predicted protein sequences, and a series of attributes describing sequence characteristics, functions, hosts, and environmental context. Users can identify plasmids of interest through attribute-based queries, comparison of plasmids within IMG/PR, or BLAST searches using user-supplied sequences.

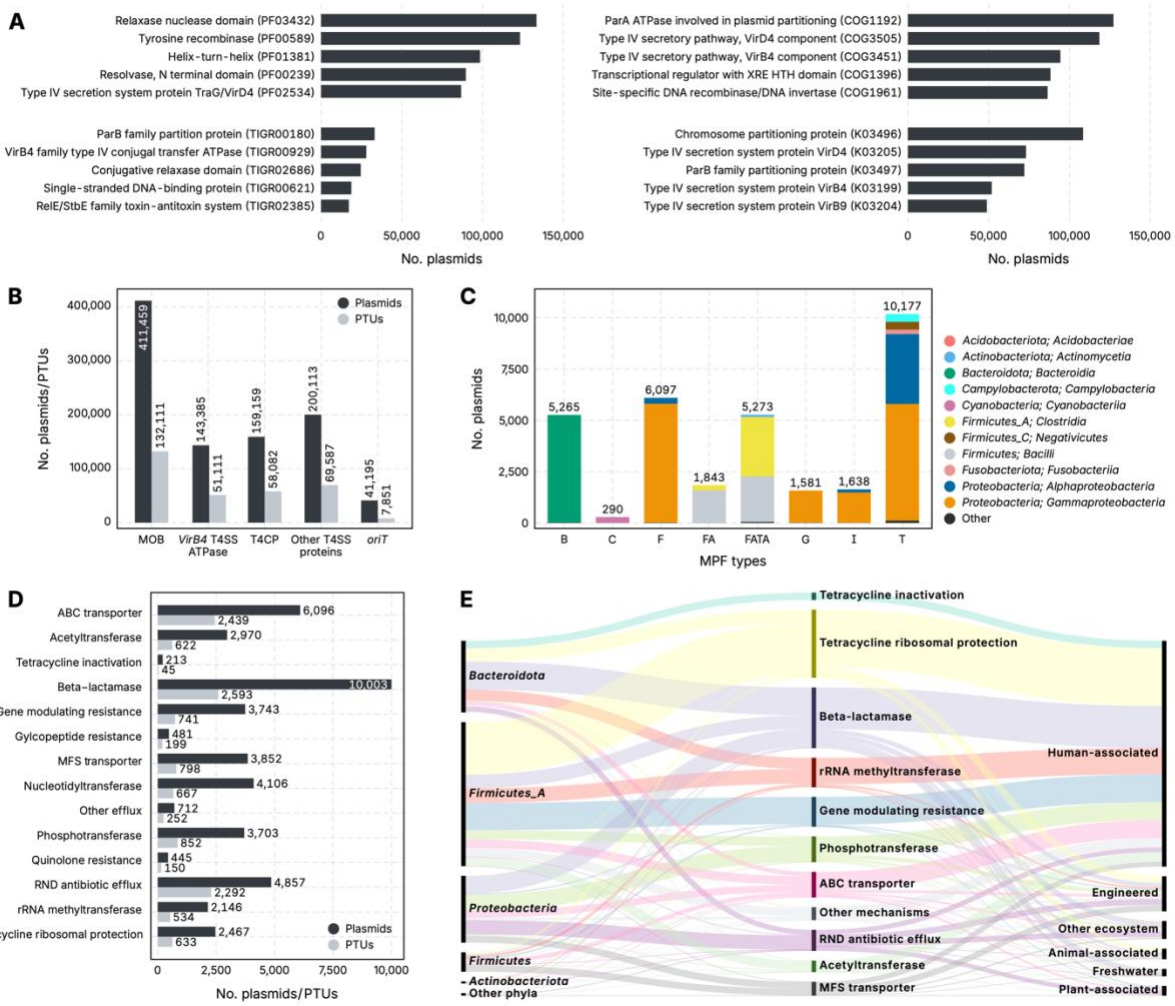
*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*


A
Source
Metagenomes (78.9%)
Isolate genomes (19.9%)
MAGs, SAGs, and metatranscriptomes (1.2%)
Putatively complete plasmids
Putatively complete (22.1%)
Fragment (77.9%)
Similarity to a reference
Similar to a reference (41.1%)
Not similar to a reference (58.9%)
Assigned host
Assigned to host taxon (39.9%)
No assigned host taxon (60.1%)
B
Length of complete plasmids (kb)
1,000
100
10
1
Isolate
Metagenome
C
PTU size
10,000
1,000
100
10
D
Intersection size
1,500
1,000
500
0
Engineered
Human-associated
Soil
Marine
Other aquatic ecosystems
Freshwater
Animal-associated
Plant-associated
Other terrestrial ecosystems
2,000
0
No. PTUs

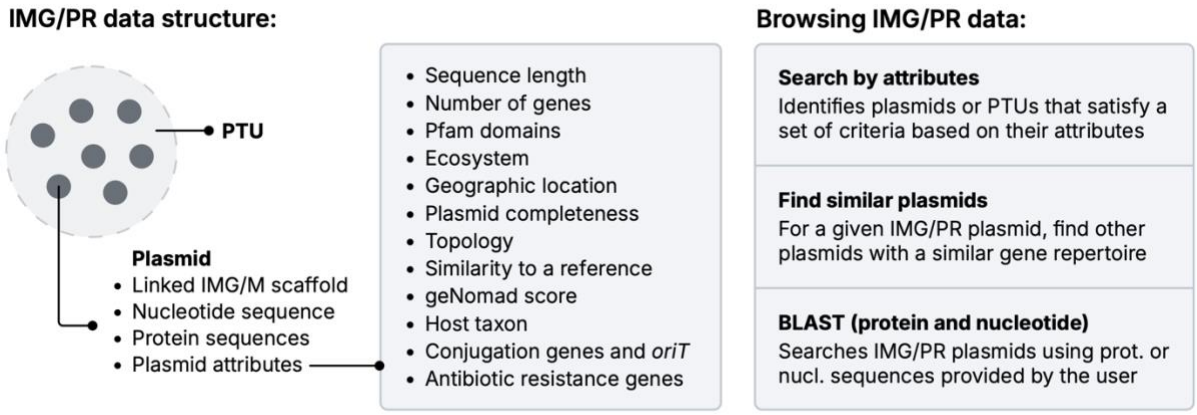



*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*


A
Bacteria
Number of sequences
Method
Proteobacteria; Zetaproteobacteria
Proteobacteria; Gammaproteobacteria
Proteobacteria; Alphaproteobacteria
Desulfobacterota; Desulfovibrionia
Myxococcota; Myxococcia
Desulfobacterota_B; Binatia
Desulfobacterota; Desulfobacteria
Desulfobacterota; Desulfobulbia
Desulfobacterota; Syntrophia
Desulfobacterota; Desulfomonilia
Desulfobacterota_F; Desulfuromonadia
Acidobacteriota; Acidobacteriae
Acidobacteriota; Blastocatellia
Bacteroidota; Bacteroidia
Bacteroidota; Rhodothermia
Bacteroidota; Chlorobia
Planctomycetota; Planctomycetia
Planctomycetota; GCA-002687715
OLB16; OLB16
Verrucomicrobiota; Verrucomicrobiae
Verrucomicrobiota; Kiritimatiellae
Chlamydiota; Chlamydiia
Elusimicrobiota; Elusimicrobia
Firmicutes_C; Negativicutes
Firmicutes_B; Desulfitobacteriia
Firmicutes_A; Clostridia
Firmicutes_A; Thermoanaerobacteria
Firmicutes; Bacilli
Chloroflexota; Ktedonobacteria
Chloroflexota; Anaerolineae
Actinobacteriota; Acidimicrobiia
Actinobacteriota; Actinomycetia
Actinobacteriota; Coriobacteriia
Deinococcota; Deinococci
Cyanobacteria; Vampirovibrionia
Cyanobacteria; Cyanobacteriia
Patescibacteria; ABY1
Patescibacteria; Saccharimonadia
Patescibacteria; JAEDAM01
Fusobacteriota; Fusobacteriia
Campylobacterota; Campylobacteria
Synergistota; Synergistia
Spirochaetota; Leptospirae
Spirochaetota; Spirochaetia
Other Bacteria
Archaea
Halobacteriota; Methanosarcinia
Halobacteriota; Halobacteria
Thermoproteota; Thermoproteia
Other Archaea
1
10
100
1,000
10,000
100,000
0%
50%
100%
Guilt-by-association
CRISPR spacer match
B
Similar to reference plasmids
Proteobacteria; Gammaproteobacteria
76.9%
Firmicutes_A; Clostridia
24.3%
Firmicutes; Bacilli
74.4%
Bacteroidota; Bacteroidia
66.8%
Proteobacteria; Alphaproteobacteria
59.7%
Other
50.2%
0%
50%
100%
C
Human-associated
(n = 54,483)
Engineered
(n = 10,732)
Animal-associated
(n = 4,141)
Freshwater
(n = 2,151)
Plant-associated
(n = 2,131)
Marine
(n = 1,668)
Actinobacteriota; Actinomycetia
Bacteroidota; Bacteroidia
Cyanobacteria; Cyanobacteriia
Deinococcota; Deinococci
Firmicutes; Bacilli
Firmicutes_A; Clostridia
Proteobacteria; Alphaproteobacteria
Proteobacteria; Gammaproteobacteria
Other




*This article has been accepted for publication by Nucleic Acids Research, published by Oxford University Press https://doi.org/10.1093/nar/gkad964*


IMG/PR data structure:
PTU
Plasmid
• Linked IMG/M scaffold
• Nucleotide sequence
• Protein sequences
• Plasmid attributes
• Sequence length
• Number of genes
• Pfam domains
• Ecosystem
• Geographic location
• Plasmid completeness
• Topology
• Similarity to a reference
• geNomad score
• Host taxon
• Conjugation genes and oriT
• Antibiotic resistance genes
Browsing IMG/PR data:
Search by attributes
Identifies plasmids or PTUs that satisfy a set of criteria based on their attributes
Find similar plasmids
For a given IMG/PR plasmid, find other plasmids with a similar gene repertoire
BLAST (protein and nucleotide)
Searches IMG/PR plasmids using prot. or nucl. sequences provided by the user