

Diverse plasmid systems and their ecology across human gut metagenomes revealed by PlasX and MobMess

In the format provided by the
authors and unedited

Supplementary Figures

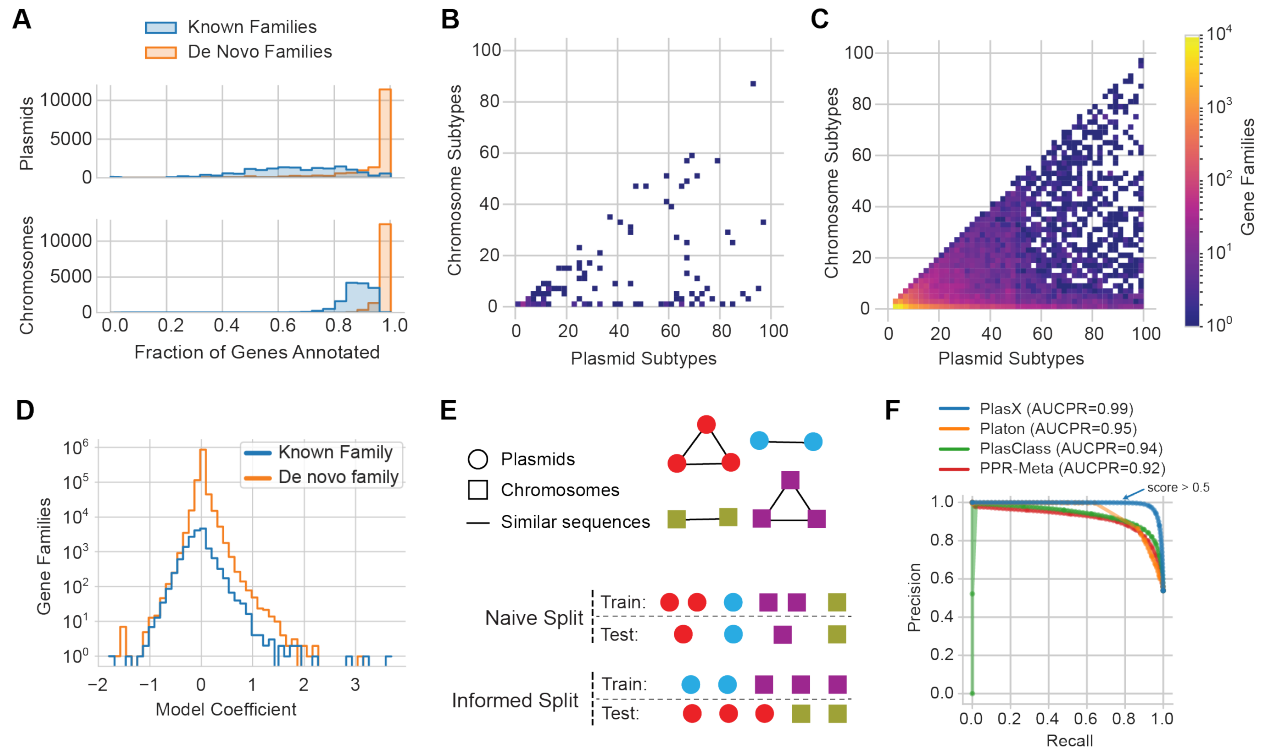


Figure S1. Additional analysis of PlasX. (A) Histograms of reference sequences, based on the fraction of genes that have known or *de novo* family annotations. (B-C) Two-dimensional histograms of known (B) and *de novo* (C) gene families, based on the number of plasmid and chromosomal subtypes that each family is found in. The number of gene families is log-scaled. Only the gene families that are enriched in plasmid subtypes (i.e. bottom-right triangular region of each plot) are shown. (D) Histograms of the coefficients learned by PlasX, showing that the vast majority of coefficients are close to zero. (E) Diagrams of different training-test split configurations for cross-validation. A random 'naive' split of plasmids and chromosomal sequences results in training and test sets that have similar sequences, due to the existence of plasmid and chromosomal subtypes that contain highly similar sequences. An 'informed' split assigns all sequences of the same subtype to either training or test, creating a more representative evaluation of a model's ability to generalize to unseen sequences. Colors and edges represent sequences that are in the same subtype. (F) Precision-recall curves using 4-fold cross-validation and a naive split.

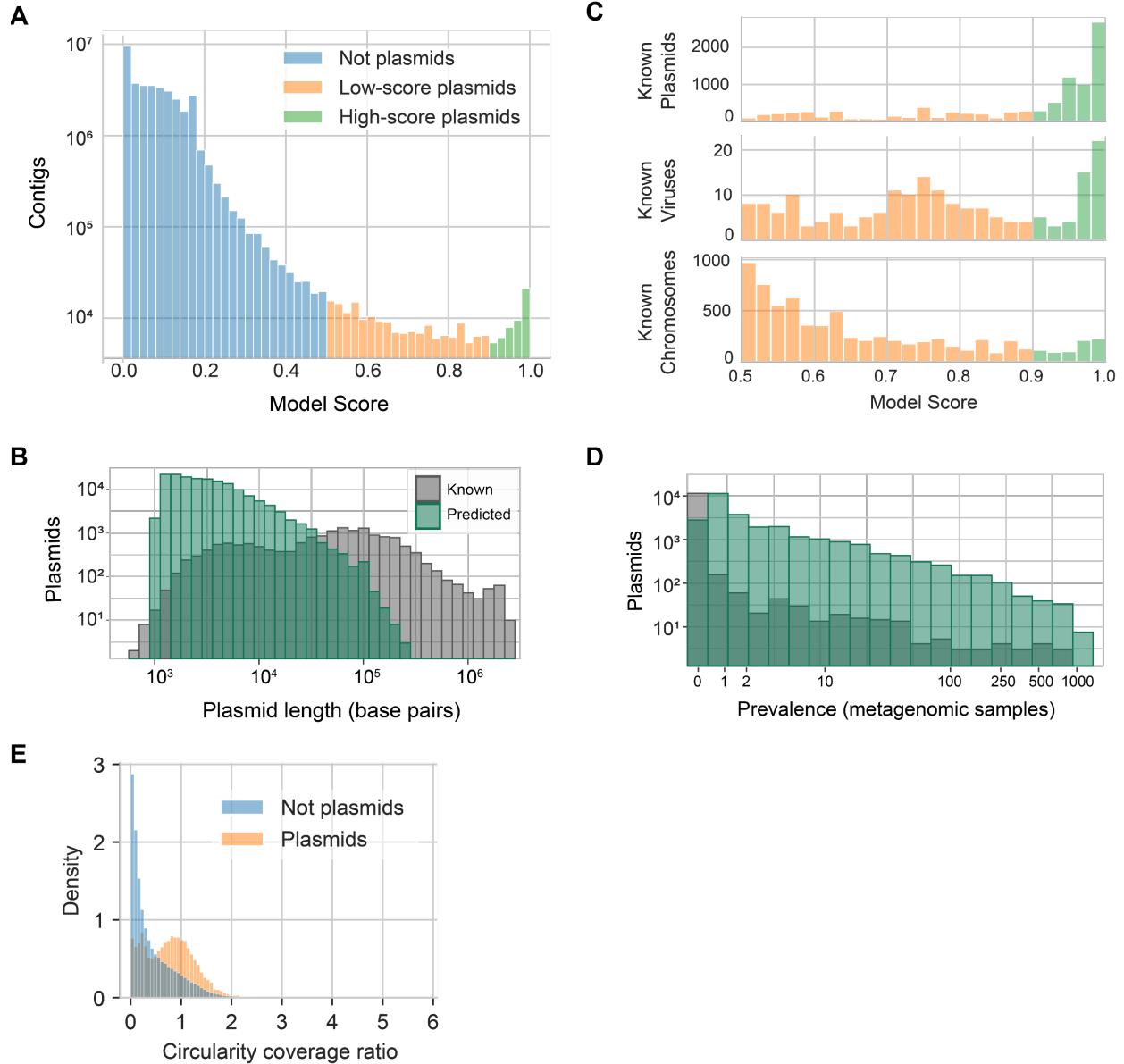


Figure S2. Additional analysis of predicted plasmids. **(A)** Model scores of all contigs assembled from all 1,782 metagenomes. 226,194 plasmids were predicted by applying a score threshold of >0.5 . Of these, 50,163 plasmids were high-scoring (≥ 0.9 score). **(B)** The sequence length of known and predicted plasmids. **(C)** Model scores of predicted plasmids that matched a sequence in NCBI ($\geq 90\%$ alignment identity and $\geq 90\%$ coverage of the predicted plasmid). Predictions are labeled as a known 'plasmid', 'virus', or 'chromosome' based on the presence of these words in the description of the matching NCBI sequence. We searched NCBI for only the filtered set of 100,719 non-fragment predictions. **(D)** The prevalence of reference and predicted plasmids across all metagenomes. **(E)** We calculated a "circularity coverage ratio" as the number of supporting reverse-forward reads divided by the average coverage of a contig. All circular contigs are shown, and they are colored if they were predicted by PlasX as plasmids with score >0.5 (orange) or not plasmids (blue).

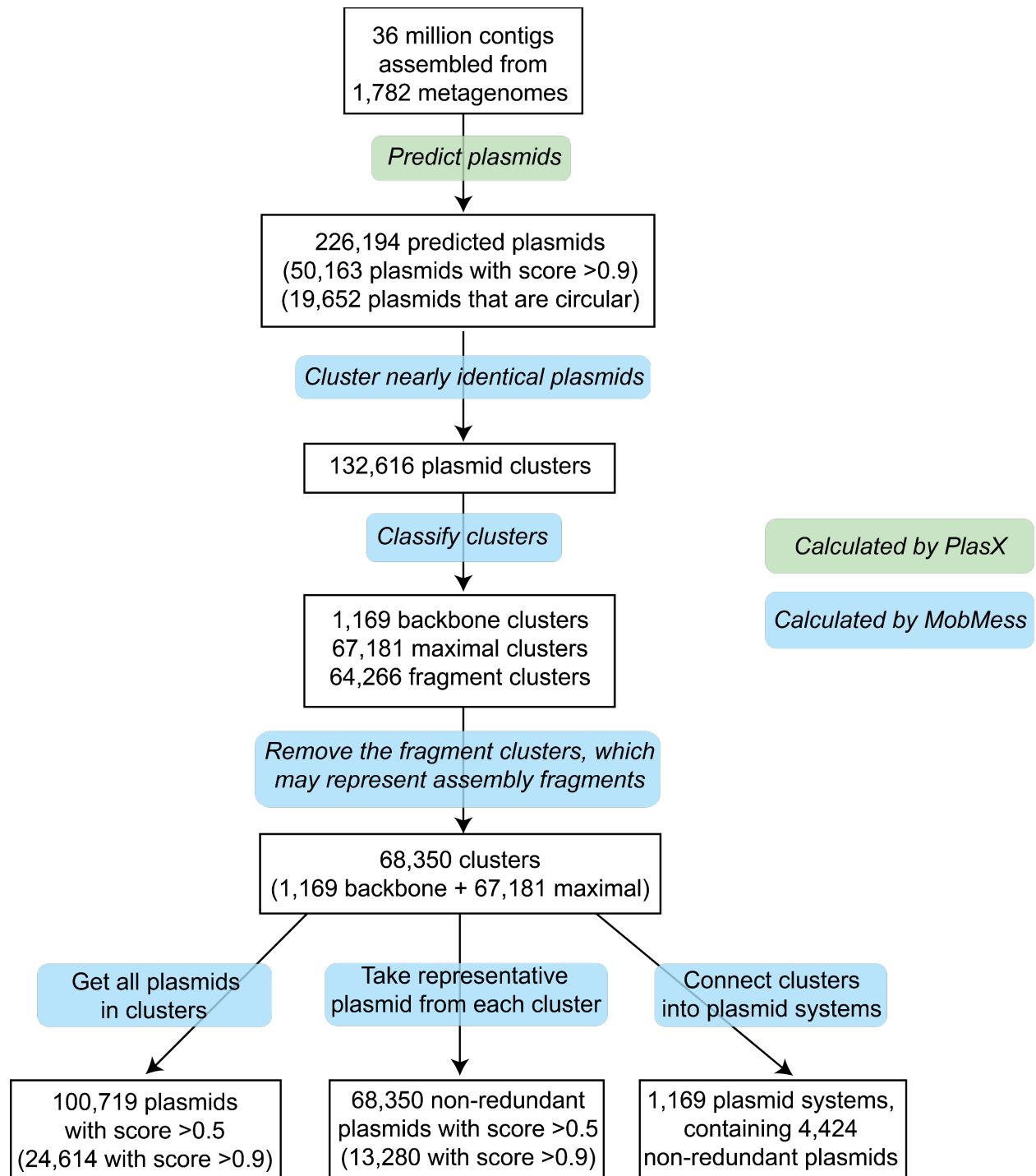


Figure S3. Workflow of predicting plasmids with PlasX and organizing them with MobMess.

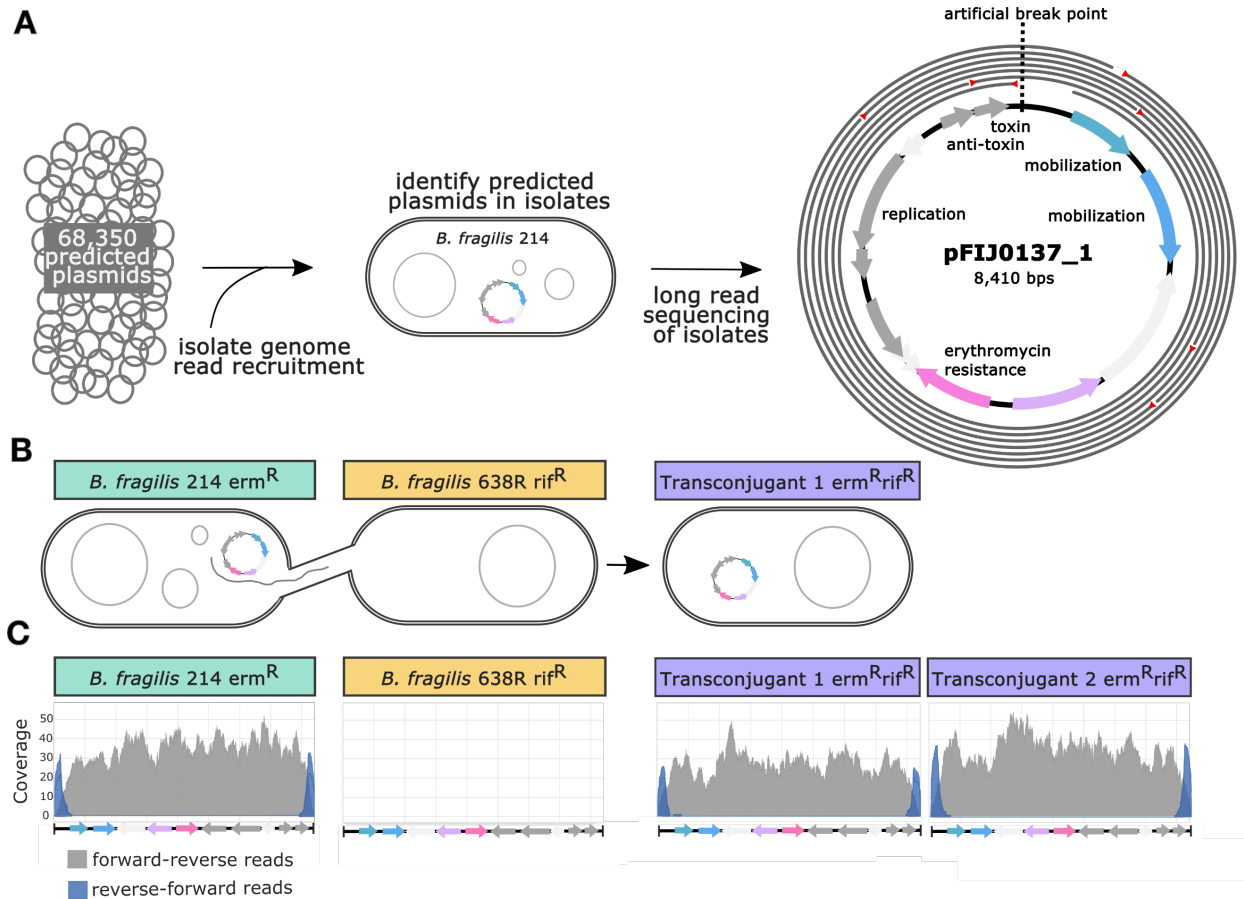


Figure S4. Experimental validation of plasmid predictions. (A) We recruited reads from the sequenced genomes of 14 *Bacteroides* isolates to determine which isolates contain our predicted plasmids. We further confirmed the presence and circularity of a predicted plasmid, pFIJ1037_1, in the isolate *B. fragilis* 214 by long read sequencing. Gray circles represent 7 (of 500) long reads that align to pFIJ1037_1. Red triangles designate the beginning of a long read. **(B)** Transfer of pFIJ1037_1 from *B. fragilis* 214 to *B. fragilis* 638R via conjugation and selection on erythromycin- and rifampicin-containing media. **(C)** Coverage plots showing read recruitment of *B. fragilis* whole-genome sequencing reads to the pFIJ1037_1 reference sequence, confirming transfer of pFIJ1037_1. Gray are forward-reverse reads, while blue are reverse-forward reads that indicate the circularity of pFIJ1037_1.

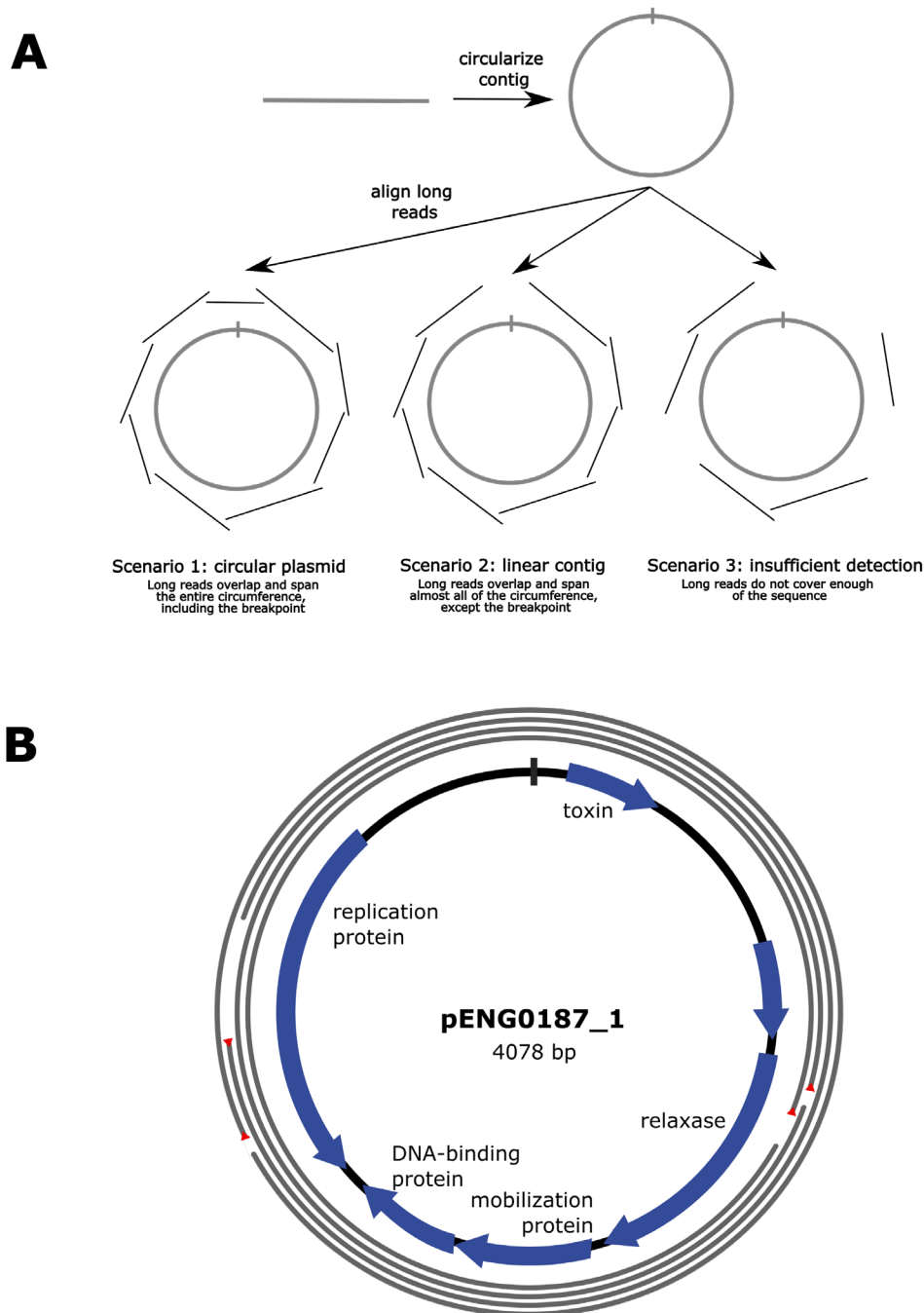


Figure S5. Long read circularity. (A) The process to identify circular plasmids using long read sequences. Contigs are always assembled as linear sequences even when originally circular in the environment. We can determine their original configuration by aligning long reads around the entire sequence. (B) Long reads from *B. fragilis* 216 aligned to pENG0187_1, demonstrating circularity. 4 of 500 reads are shown for simplicity. Red triangles designate the beginning of a long read.

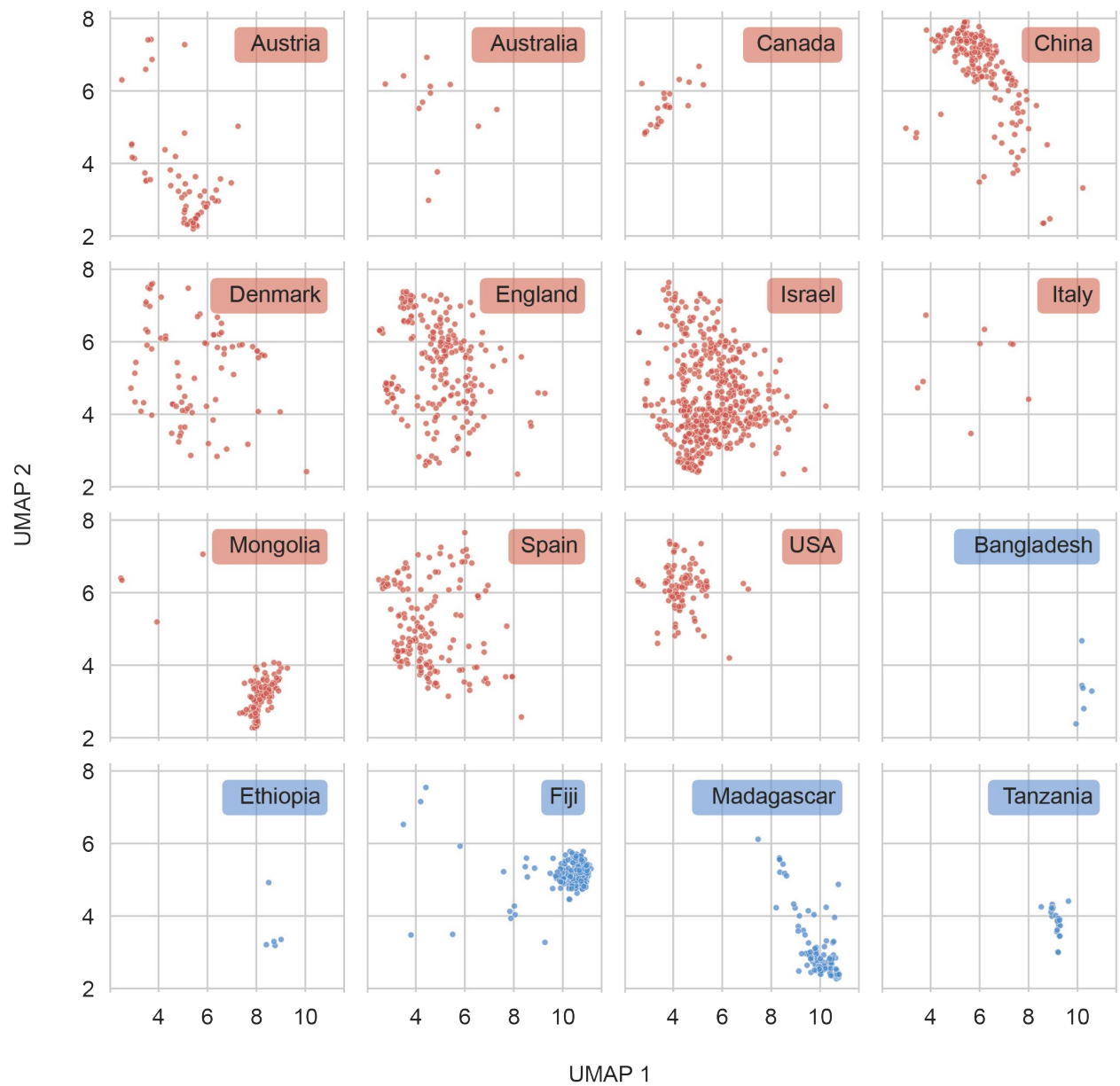


Figure S6. UMAP plot as in Figure 3C. Metagenomes have been partitioned to show clustering within each country.

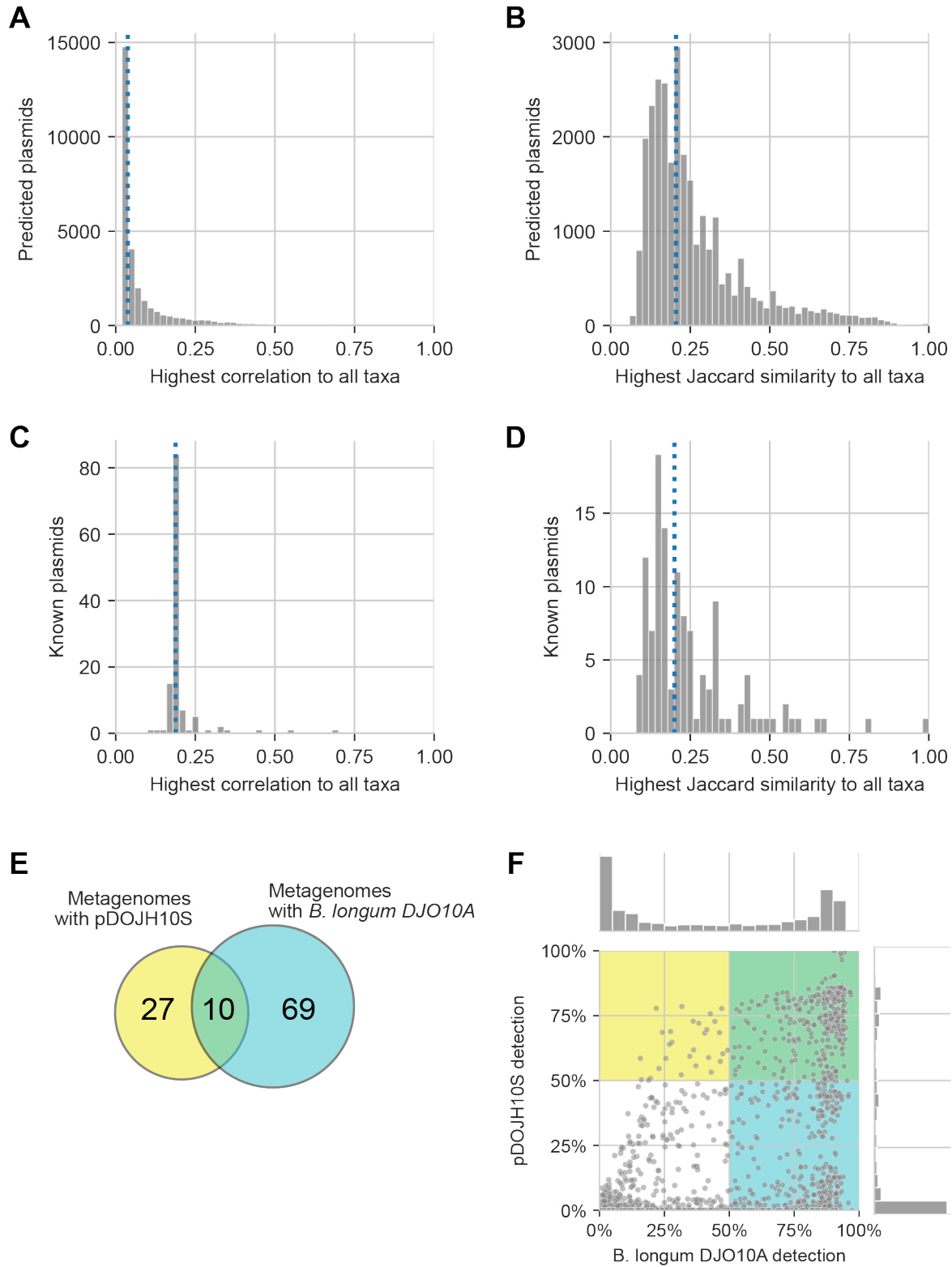
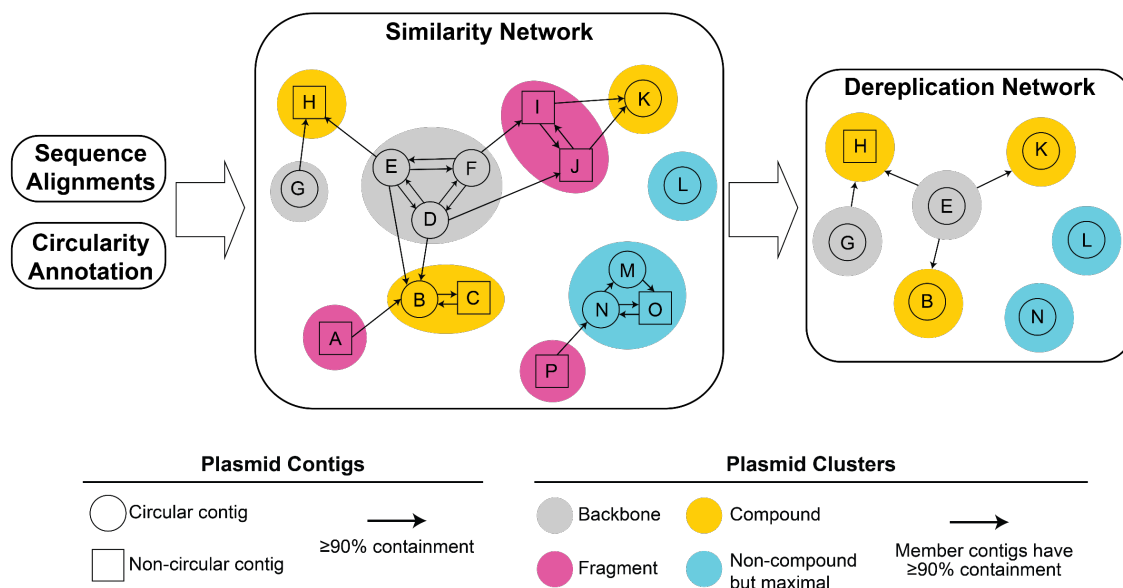


Figure S7. Comparison of the ecological distributions of plasmids and microbial taxonomy. We measured the association between every plasmid and taxon by calculating the correlation between their abundance levels across metagenomes, using the SparCC technique¹. As another association measure,

we applied thresholds to the abundance levels and then calculated the Jaccard similarity between the metagenomes containing the plasmid versus those containing the taxon. We estimated taxon abundances with bracken². We restricted analyses to plasmids that were present in at least 5 metagenomes. **(A-B)** For every predicted plasmid, we identified the taxon with the highest correlation **(A)** or Jaccard similarity **(B)**. **(C-D)** We did the same to identify the best matching taxa of reference plasmids. Blue lines indicate the median of each distribution. **(E)** Venn diagram showing the discordance between the metagenomes containing a plasmid pDOJH10S and those containing its cognate host, a *B. longum* strain. **(F)** Detection of pDOJH10S and the *B. longum* strain, based on read recruitment instead of bracken. Each point represents a metagenome. Yellow and blue rectangles highlight the metagenomes where the plasmid or strain, respectively, are identified as being present with >50% detection.

A



B

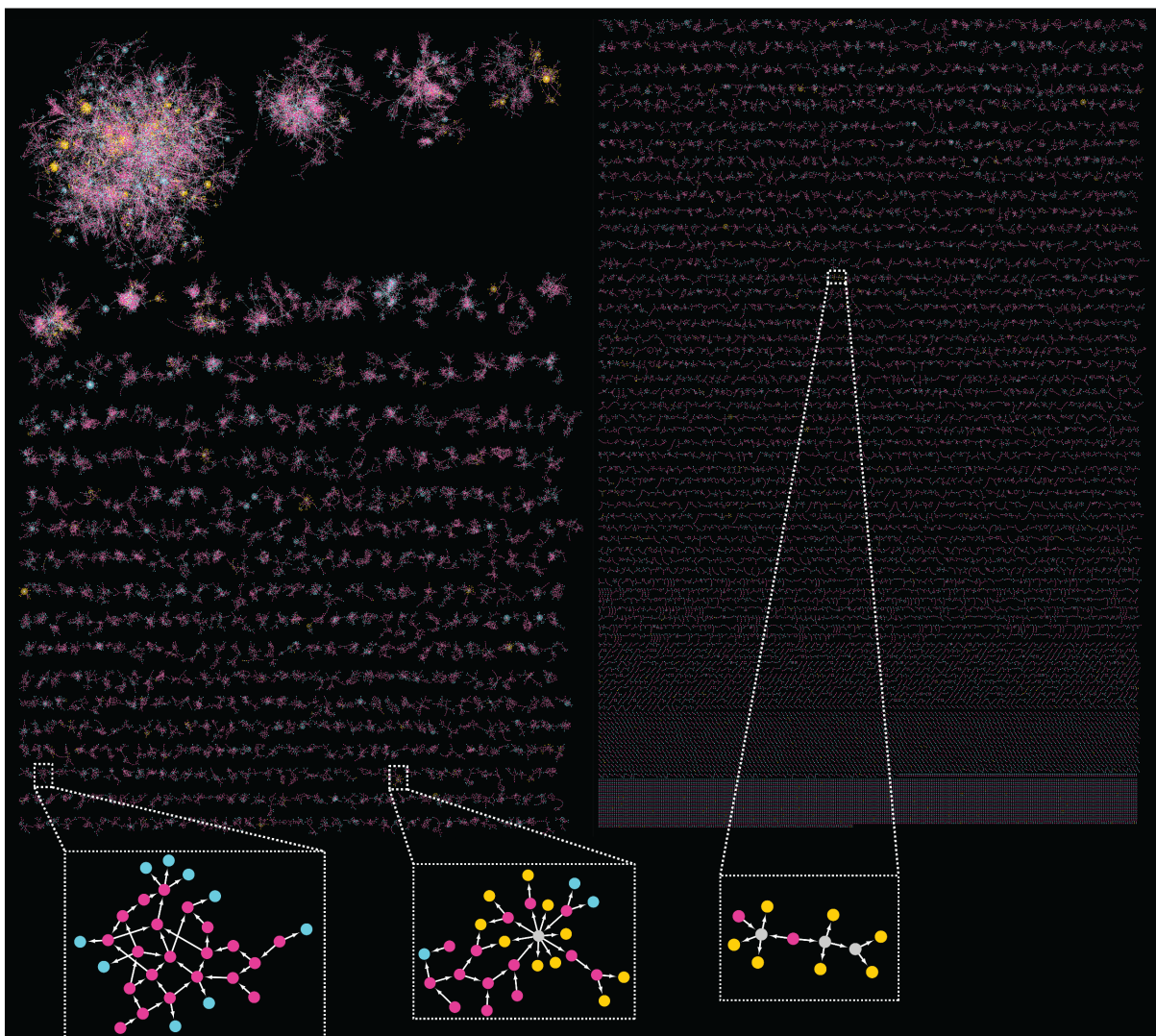


Figure S8. The MobMess algorithm and application to predicted plasmids. (A) Diagram of the MobMess algorithm for dereplicating plasmids and discovering plasmid systems. All-vs-all sequence alignments and circularity information are used to construct a similarity network of plasmid contigs. Similar contigs are clustered, and every cluster is labeled as either a backbone, fragment, compound, or non-compound maximal. A plasmid system consists of a backbone cluster and the compound clusters connected to the backbone. This example shows two systems: one system has G as the backbone (H is the compound plasmid), and another system has D, E, and F as the backbone (B, C, H, and K are the compound plasmids). To dereplicate, fragment clusters are discarded and a representative sequence is chosen for every non-fragment cluster. **(B)** Network of clusters of predicted plasmids. All clusters are shown except those that are not connected to any other cluster.

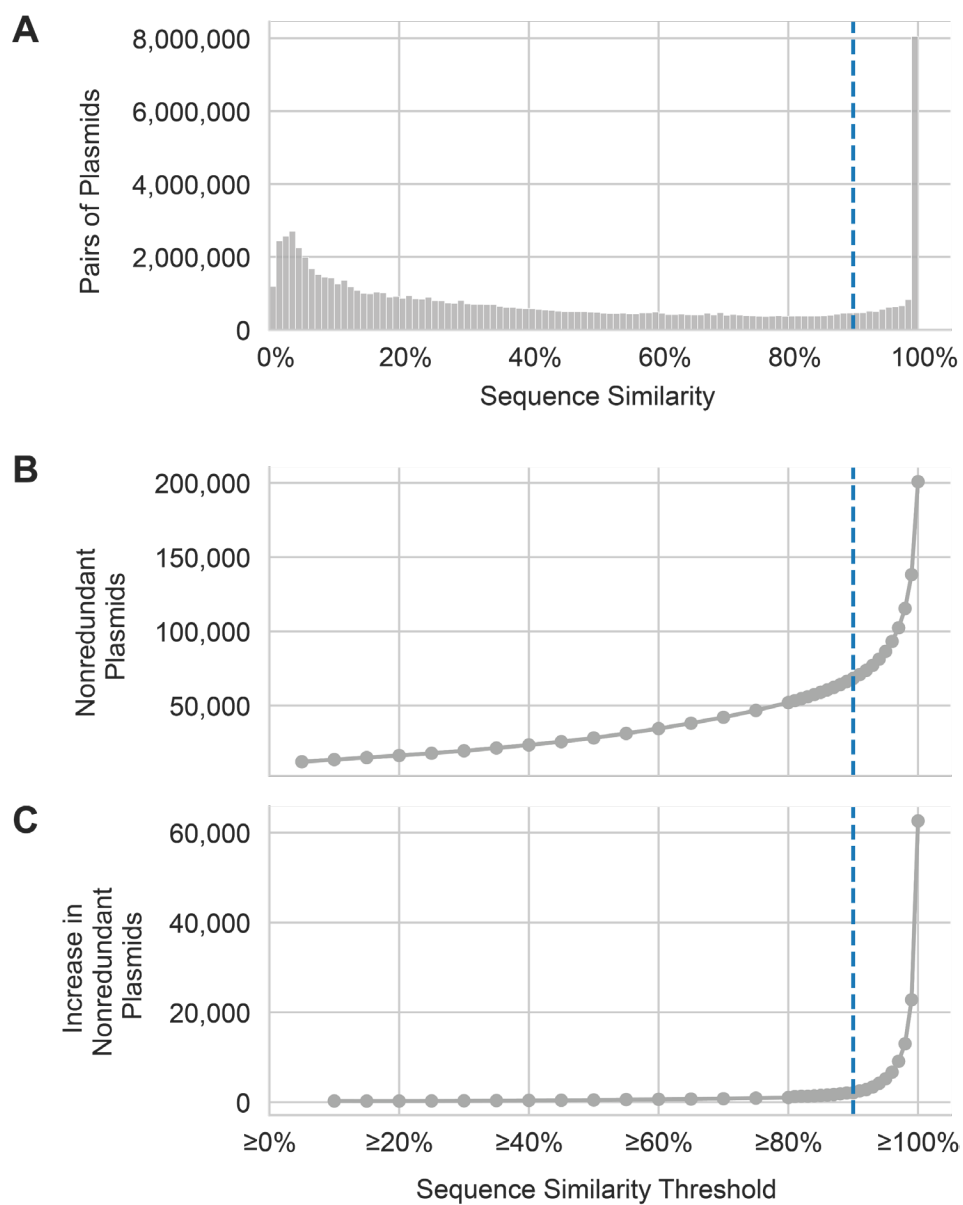


Figure S9. Choosing a similarity threshold for MobMess. (A) Histogram of similarities between every pair of the 226,194 predicted plasmid contigs. (B) We ran MobMess using different thresholds on the similarity, and then we calculated the number of non-redundant plasmids generated. (C) The derivative of the curve in B. The blue dashed lines represent our current $\geq 90\%$ similarity threshold.

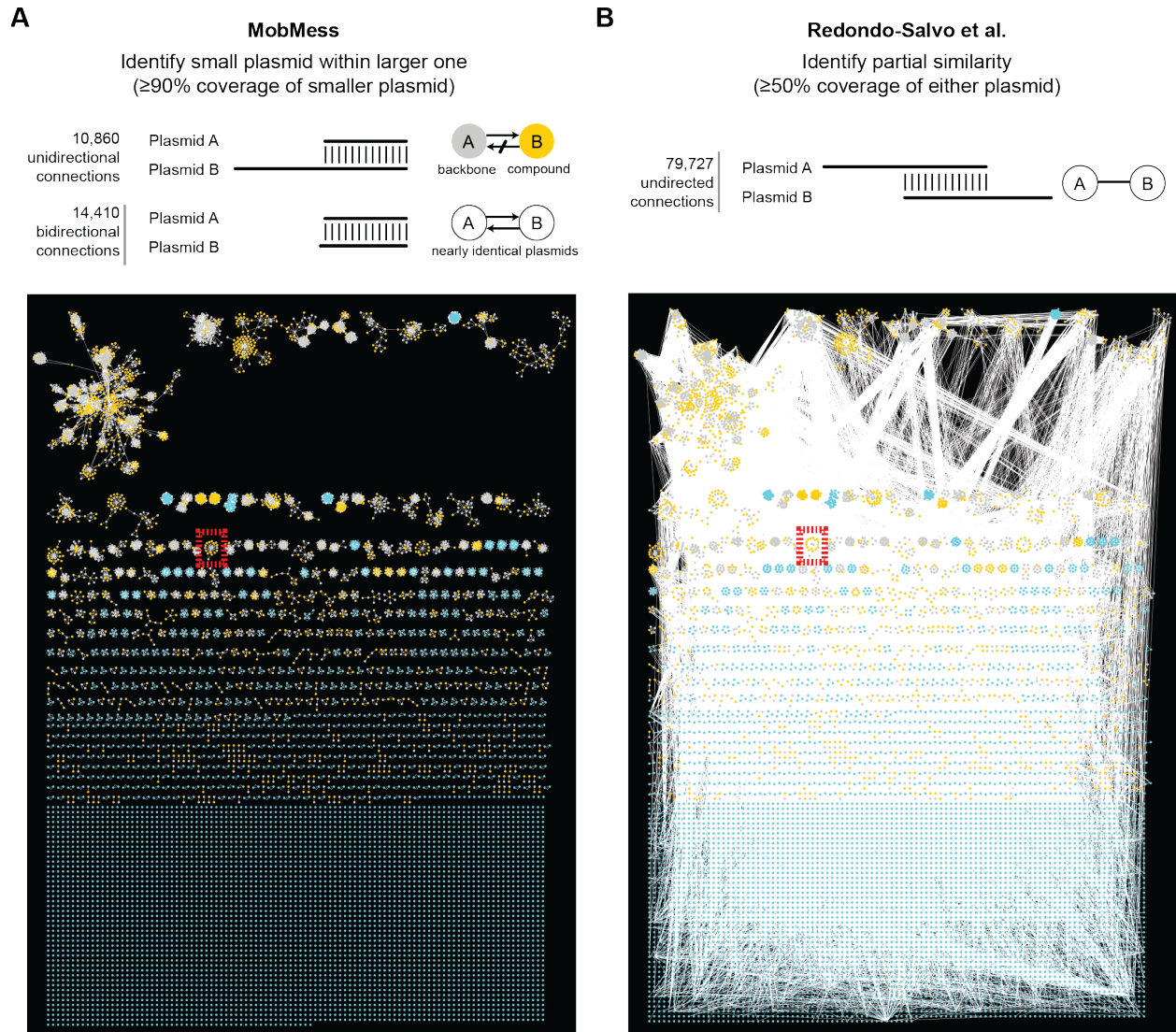


Figure S10. Conceptual differences in constructing plasmid similarity networks. We ran MobMess on the set of 9,894 reference plasmids analyzed by Redondo-Salvo et al.³. MobMess constructs a network with directed edges, by aligning plasmids and determining if one plasmid is found as a subsequence within another. Redondo-Salvo et al. constructs a network with undirected edges, by determining whether two plasmids contain partial homology. **(A-B)** Visualization of the similarity networks. We used Cytoscape⁴ and the Prefuse directed layout algorithm⁵ to lay out the nodes in the MobMess network (A), and then we applied the same layout to the Redondo-Salvo et al. network (B). The red boxes represent the example shown in Figure S11.

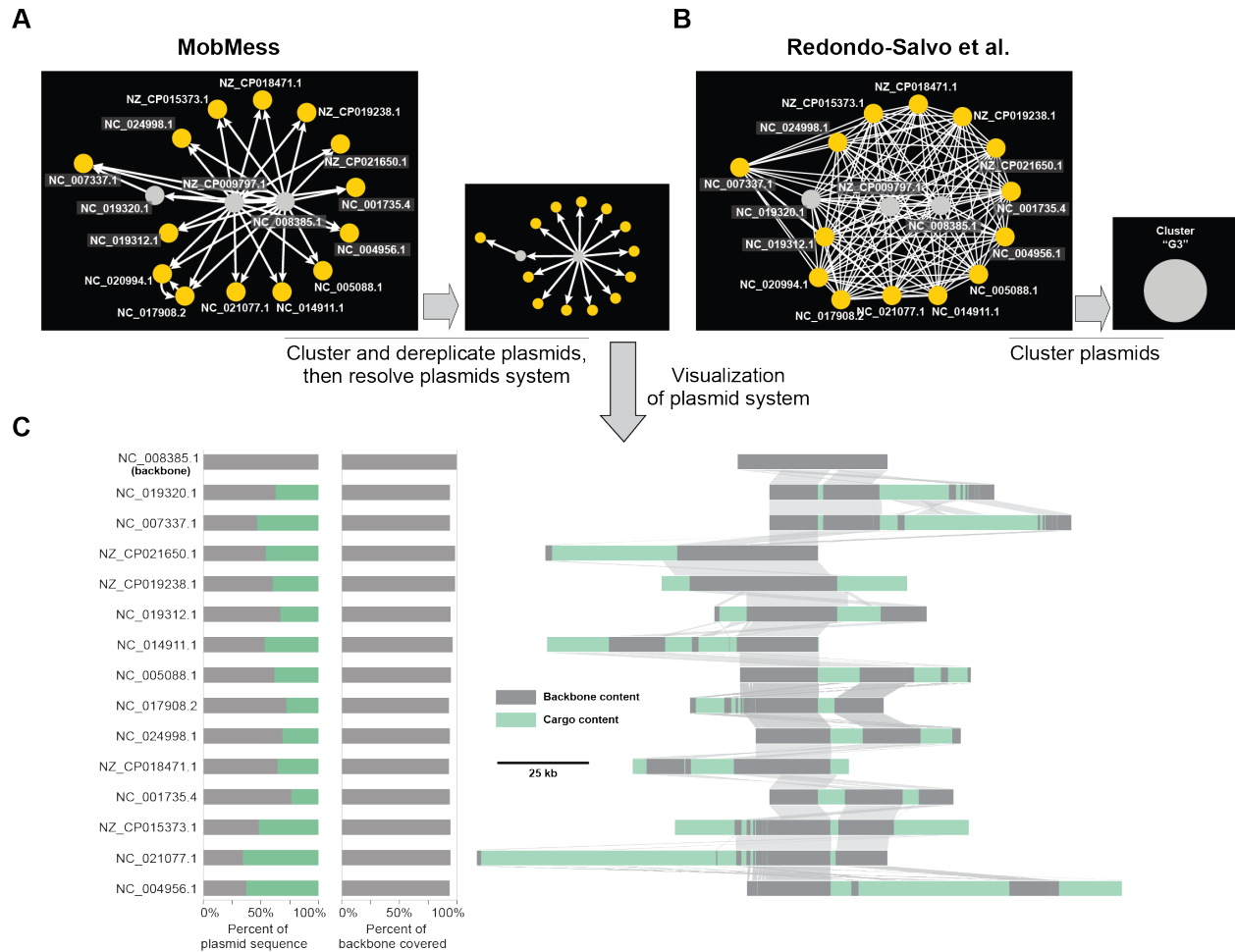


Figure S11. Comparison of MobMess versus Redondo-Salvo et al.³ for studying a plasmid system. (A-B) An example from the similarity networks in Figure S10, showing the connections between 17 plasmids from the same plasmid system. MobMess further collapses its network to dereplicate plasmids and reveal the plasmid systems's "star"-like topology, where a backbone connects to its compound plasmids. Redondo-Salvo et al. did recognize that these plasmids are related (represented by a cluster called "G3"), but they connected almost every pair of these plasmids in a "hairball" topology, obfuscating the system's internal organization. (C) Alignments of plasmids in the MobMess system. Subregions in every sequence are colored gray or green to represent backbone or cargo content, respectively. Ribbons between sequences represent the alignment of subregions. The barcharts show the total breakdown of each plasmid into backbone versus cargo, as well as the fraction of the backbone sequence ("NC_008385.1") that is found within the plasmid.

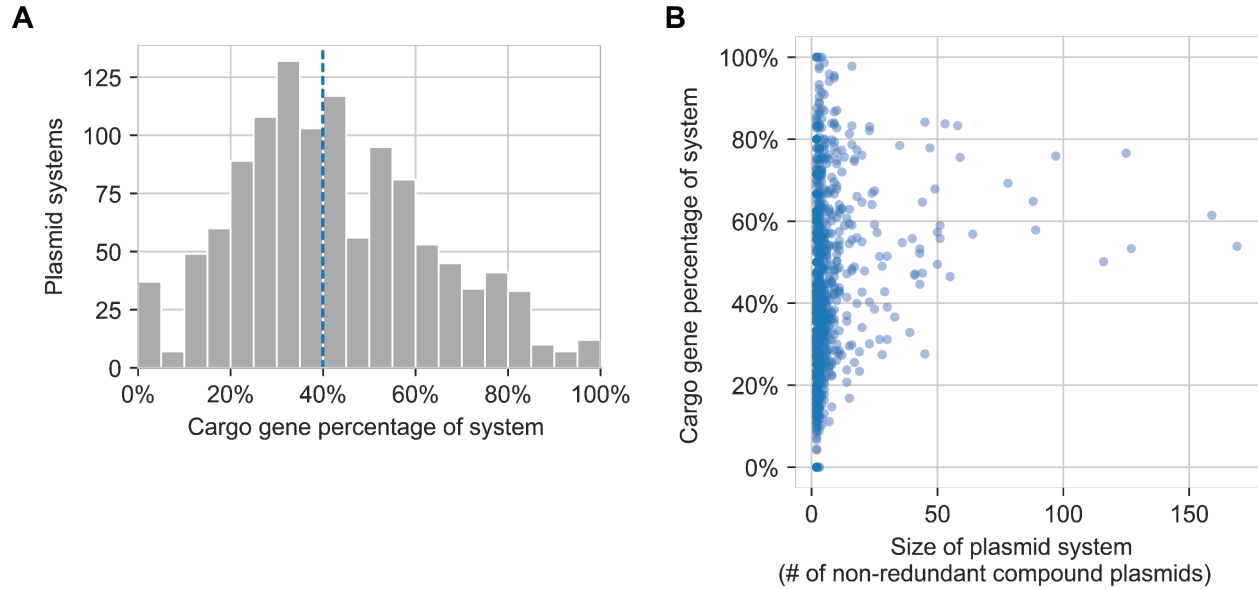


Figure S12. Backbone and cargo composition of plasmid systems. (A) For every plasmid system and compound plasmid in the system, we calculated the percentage of genes on the compound plasmid that were classified as cargo versus backbone genes (see Methods). We then averaged the cargo gene percentages across all compound plasmids in the system (x-axis). The vertical blue line shows the median at 40%. **(B)** Scatterplot of the cargo gene percentage versus the size of a plasmid system, showing a lack of correlation ($R^2 = 0.03$). We defined the size as the number of non-redundant compound plasmids.

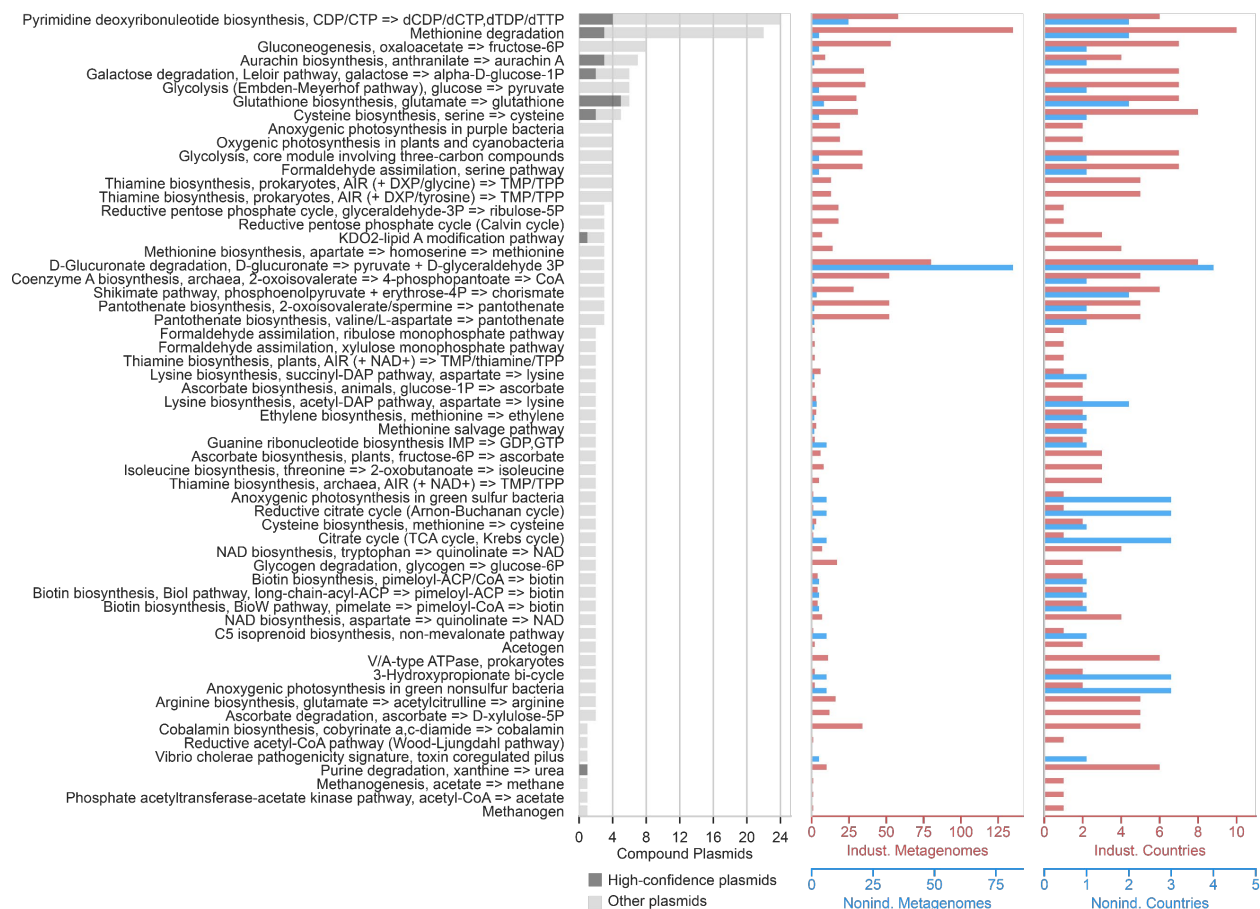


Figure S13. Functional annotation of cargo genes to KEGG modules, similar to Figures 5A and 5B. This plot excludes KEGG modules that occur in only one plasmid system. To avoid redundancy with Figure 5A, this plot also excludes modules that occur in cargo genes annotated to antibiotic resistance.

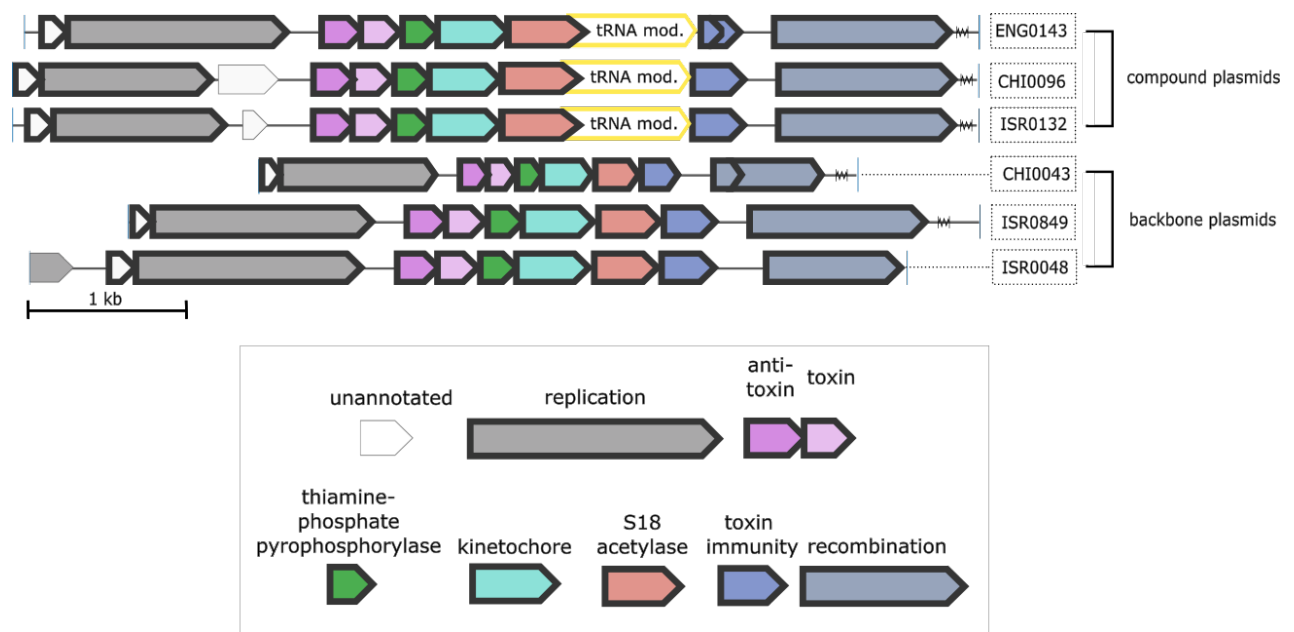


Figure S14. Plasmid system PS1110. Compound plasmids in this system contain a gene that encodes two enzymes, a tRNA Gm18 2'-O-methylase (yellow, 'tRNA mod.') and a Ribosomal protein S18 acetylase (red). Backbone plasmids contain a similar gene that encodes the S18 acetylase but lacks the tRNA methylase. Backbone genes have a thick, black outline.

Supplementary Notes

Publicly Available Human Gut Metagenomes

We downloaded FASTQ files for 1,782 short-read and paired-end gut metagenomes from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) using the program `fastq-dump`. The countries represented in our collection were Austria⁶, Australia (<https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092>), Bangladesh⁷, Canada⁸, China^{9,10}, Denmark¹¹, England¹², Ethiopia¹³, Fiji¹⁴, Israel¹⁵, Italy¹⁶, Madagascar¹³, Mongolia¹⁷, Spain¹⁸, Tanzania¹⁶ and USA^{19,20}. Some samples were sequenced multiple times (i.e. multiple records in SRA), in which case we concatenated the FASTQ files together. While building our search terms, we have separated these multiple accessions using the delimiter '|

Relevance of PLSDB database for predicting gut plasmids

By training PlasX on all known plasmids in the 2019_03_05 version of the PLSDB database, we designed it to be suitable for identifying plasmids in metagenomes from any environment. The vast majority of plasmids in PLSDB, however, originate from aerobic organisms, with a relatively lesser representation of anaerobic taxa. Our examination of the microbial taxa and isolation sources of the plasmids in PLSDB suggested that the 23% (3772/16168) of plasmids in this database originated from the phylum Bacteroidetes or Firmicutes, each of which includes many organisms that are commonly found in anaerobic human gut environment. In addition, 6% (884) of the plasmids were directly annotated as being isolated from a gut sample. Together, 28% (4536) of plasmids were either isolated from an organism that resolved to Bacteroidetes or Firmicutes, or they were isolated directly from a human gut sample. This unevenness in PLSDB reflects a common limitation of any database, where some types of sequences are more represented than others. To minimize the impact of any single type of sequence from dominating PlasX's logic, and thereby enable it to pay attention to all types of sequences, we explicitly grouped the training sets of plasmid and chromosomal sequences into subtypes and then assigned weights to individual sequences in such a way that every subtype had an equal total weight in the model (Methods). Nevertheless, in theory, the relatively low representation of plasmids from anaerobic organisms in the training dataset can make PlasX more prone to miss true plasmids ("false negatives") when applied to anaerobic environments such as the human gut microbiome, compared to its applications to environments that are aerobic. Despite the risk of

increased proportion of false negatives, PlasX was still able to predict a large number of plasmids from human gut metagenomes. PlasX was also able to classify pWCP as a plasmid (score=0.73), which is a recently characterized plasmid of *Wolbachia* (a likely anaerobe as described by Fallon et al.²¹ and Uribe-Alvarez et al.²²) that was missed by other modern plasmid prediction algorithms.

Additional validations of predicted plasmids

To determine if a predicted plasmid has canonical plasmid features, we ran MOB-suite²³. This tool searches a sequence for known examples of four types of features: plasmid replicon (e.g. replication genes), relaxase, mating pair formation, and origin of transfer. We installed MOB-suite v3.0.1 using pip, in an Anaconda Python environment that has mash v2.2. We ran the MOB_typer subroutine (command ``mob_typer``) using default parameters and followed the execution instructions at <https://github.com/phac-nml/mob-suite>, and summarized the results in Table S8.

To determine if a predictive plasmid was previously characterized, plasmids were blasted against NCBI using the blast package (v2.9.0, installed via bioconda). On October 13, 2021, we downloaded version 5 of the NCBI databases non-redundant nucleotide (nt), ref_prok_rep_genomes, ref_viroids_rep_genomes, and ref_viruses_rep_genomes, and then integrated them into a single database using the ``blastdb_aliastool`` command. We then searched every predicted plasmids against this combined database, using the ``blastn`` tool with the ``-task megablast`` parameter for efficient searching. For each plasmid, we examined all matching NCBI sequences (called 'subjects') and chose the one with the highest 'qcovs' (query coverage per subject), which represents the fraction of the plasmid sequence that is covered by all high-scoring segment pairs (HSP). Tiebreaking was done by sorting subjects by the maximum bitscore of the HSPs. If the qcovs of the best matching sequence was $\geq 90\%$, then we considered the predicted plasmid as found in NCBI and further categorized the matching sequence by searching for the keywords 'plasmid', 'virus', 'chromosome' (in that order, disregarding capitalization) in its NCBI description. For example, if the description of the matching sequence contained the word 'plasmid', then we said the predicted plasmid matched a known plasmid on NCBI. Similarly, if the description contained 'chromosome' but not 'plasmid' nor 'virus', then we said that the predicted plasmid matched a known chromosome on NCBI. If the qcovs of the best sequence was $< 90\%$, then we labeled the predicted plasmid as not found in NCBI.

We further investigated the subset of predictions that were highly similar to a sequence in NCBI and categorized matches as either known plasmids (26.9%), chromosomes (21.3%), viruses (0.6%), or an unclear type of sequence (51.2%) (Figure S2C). A total of 189 predictions matched a known virus. Of these, 110 were recognized as plasmids by MOB-suite or keywords but also contained virus-related COG or Pfam functions, as indicated by the keywords 'virus', 'viral', and 'phage'. These predictions carry both plasmid and viral features, a phenomenon that has previously been reported^{24–27}. Surprisingly, 808 predictions that matched a known chromosome were also circular-associated and recognized by MOB-suite or plasmid keywords. One explanation of these data is that these plasmids can switch between an extrachromosomal or a chromosome-integrated state.

We also found that among all of the assembled contigs that were circular, those that were predicted as plasmids tended to have a higher ‘circularity coverage ratio’, defined as the number of reverse-forward read pairs supporting circularity by the read coverage (Figure S2E). Thresholding this ratio could be used as an additional filter in future work to identify plasmids of higher confidence.

Comparison of PlasX to other tools

Here we implemented a more realistic evaluation framework to compare PlasX to three state-of-the-art algorithms, PlasClass²⁸, PPR-Meta²⁹, and Platon³⁰. We first evaluated performance in 4-fold cross-validation, using a ‘naive’ randomized splitting of sequences into training and test data (Figure S1E). PlasX achieved nearly perfect accuracy, with the highest area under the precision-recall curve (AUCPR=0.99) compared to all other methods (Figure S1F). While naive splitting is a common evaluation technique, it is not a fair strategy as it can separate very similar sequences into training and test data, especially given the redundancy of sequences in public databases, and thus inflate the accuracy of classification. As a more accurate benchmark, we (1) designed an ‘informed’ split by first clustering plasmid and chromosomal sequences into subtypes and then keeping all sequences in the same subtype together in either the training or test data to better evaluate the ability to recognize sequences that are distantly related to the training data and (2) assigned normalized weights to sequences to prevent well-studied plasmids from influencing the prediction ability disproportionately (see Methods). This advanced benchmark revealed a greater performance divide between PlasX (weighted AUCPR=0.70) and all other methods, with the next best method performing substantially worse (Platon, weighted AUCPR=0.23) (Figure 1D).

Additional notes on the execution of other plasmid prediction tools and benchmarks

The purpose of this section is to share installation and runtime details of our benchmarking of PlasX against state-of-the-art tools, Platon³⁰, PlasClass²⁸, PPR-Meta²⁹, and Deeplasmid³¹, and our use of the publicly available data for this aim.

We downloaded PlasClass from <https://github.com/Shamir-Lab/PlasClass> (v0.1.0-2-gb80a4f4). We downloaded PPR-Meta from <https://github.com/zhenchengfang/PPR-Meta> (v1.0-14-gab99c91). We downloaded Platon from <https://github.com/oschwengers/platon>, and then modified the code to more efficiently parallelize across many CPUs (modifications at <https://github.com/michaelkyu/platon>).

To ensure a fair comparison of models in cross-validation (Figures 1D and S1F), we did not use the pretrained versions of PlasClass and Platon from their original studies, but instead we retrained those models using the same training sequences as we used for PlasX in each cross-validation fold. We retrained PlasClass on the 10 kbp slices in each fold. For computational feasibility, we retrained Platon on the whole sequences in each fold, instead of slices. We did not train PlasClass and Platon with sequence weights because they don't take in weights as input, but we did calculate precision and recall with weights. We used Platon's RDS score as its final prediction score, ignoring whether it found other features like conjugation and replication genes, as Platon was unable to identify them in a feasible runtime when evaluating all 10 kbp slices. PPR-Meta and Deeplasmid do not provide software interfaces for retraining models on specific datasets, so we ran the pretrained versions of these models from their original studies (note that those studies used different sequence datasets).

We downloaded the four sequence versions of the Wolbachia plasmid pWCP from <https://doi.org/10.6084/m9.figshare.6380015> (Table S8). We made predictions of pWCP using the original published model versions of PlasClass, Platon, PPR-Meta, and Deeplasmid.

We also downloaded the more recent 2021_06_23_v2 version of PLSDB, which was released after we trained PlasX. This version contains 34,513 plasmid sequences, of which 21,012 were not used to train PlasX, so we focused our evaluation on this subset of plasmids to measure PlasX's generalizability (Table S2). As a comparison, we ran the original pretrained version of Platon in all execution modes ('sensitivity', 'specificity', 'accuracy', and 'characterize').

We downloaded the collection of all ICE sequences (n=552) from ICEberg³² 2.0 at <https://db-mml.sjtu.edu.cn/ICEberg/> on September 30, 2022. We also downloaded 455 prophage sequences from the NCBI Virus data portal (<https://www.ncbi.nlm.nih.gov/labs/virus>) on September 30, 2022. To download them, we selected the “Bacteriophages” subset from the “>Find Data” menu bar, and then we applied filters of “Only” for the “Provirus” option and “complete” for the “Nucleotide Completeness” option. We made predictions of these ICE’s and virus sequences using the original pretrained version of Platon, using its default ‘accuracy’ mode (Tables S4 and S5).

For the four external datasets described above (pWCP, PLSDB’s recently added plasmids, ICE’s, and viruses), we ran PlasX and other plasmid tools on the whole sequences (instead of 10 kbp slices).

We ran Deeplasmid using the Docker image of the CPU implementation, following instructions at <https://github.com/wandreopoulos/deeplasmid> (version sha256:10809927e2c8a14cf86231801b804b0bd4bddf600821d17fd8b7e41a15c562c0). While we were able to run Deeplasmid on the Wolbachia plasmid pWCP, it was prohibitively slow to run on the entire set of 10 kbp slices used for cross-validation evaluation. In particular, we found that Deeplasmid running on a MacOS laptop takes ~3 hours for 1,000 slices, so we estimated it would take ~3.7 years to run on all slices. While the GPU implementation of Deeplasmid might be able to run faster, we were unable to execute its prebuilt Docker image (version sha256:f3a22993fb765a7f9678b174245b64976e7e52a4dce85570060900b794af5e43). We suspect that this image is incompatible with modern machine setups, like ours, because Deeplasmid depends on software that is several years old. For example, it requires the CNTK library, for which development was abandoned over 3 years ago (https://docs.microsoft.com/en-us/cognitive-toolkit/releasenotes/cntk_2_7_release_notes). We were also unable to rebuild a Docker image to run the GPU implementation, despite attempts to modify the Docker build file (see the issue we raised at <https://github.com/wandreopoulos/deeplasmid/issues/3>).

Comparison of MobMess to other plasmid clustering methods

To design MobMess, we first examined the histogram of similarities between all pairs of predicted plasmids, revealing an average nucleotide identity (ANI) “valley” with the lowest point at around 85-90% identity (Figure S9A). However, this valley was wide and shallow, reflecting the occurrence of many plasmids that share partial similarity between ~20% to 90% identity. This

shallow valley could have emerged due to a number of reasons, such as assemblies from different metagenomes containing sequence fragments that are partially redundant with each other. Another explanation is the dynamism of plasmids—its tendency to recombine with other plasmids, other mobile elements, or chromosomes—resulting in a mosaic composition of genetic material that originated from different sources³³. In contrast, a deeper valley has been observed when analyzing collections of reference plasmids^{3,34} and of bacterial taxa^{35,36}. We believe this reflects reference databases containing non-redundant genomes from distant branches of life, such that clear boundaries exist when grouping those genomes into evolutionary clusters. Consequently, previous methods for clustering reference plasmids chose an identity threshold in the middle (e.g. >50% ANI by Redondo-Salvo et al.), as slightly shifting the threshold up or down would have minor effects on clustering. The valley in our collection of metagenome-derived plasmids appeared to lack a clear ANI threshold, as every threshold within the range of 20% to 90% seemed to be almost equally reasonable targets (Figure S9A). Choosing a single threshold appeared to over-split or over-combine plasmids, rather than define ecologically and/or evolutionarily cohesive units. Thus, we designed MobMess not to simply rely on an ANI threshold but, in addition, take a more nuanced examination of the topology of the sequence similarity network by considering how much one sequence is contained within another sequence.

To identify an appropriate threshold on sequence containment (I_{local} and C , defined in Methods), we examined MobMess's behavior across a wide range of thresholds. As the threshold is made stricter, MobMess gradually separated plasmids into distinct clusters, and consequently the number of non-redundant plasmids increased (Figure S9B-C). This growth in non-redundant plasmids occurred at a mostly constant rate from a threshold of 10% to 90%, but it suddenly accelerated from 90% to 100%. These results suggest that a threshold stricter than $\geq 90\%$ (e.g. $\geq 95\%$ or $\geq 99\%$) would split highly similar plasmids into separate clusters. Thus, we found that $\geq 90\%$ alignment identity and coverage was a natural threshold to define containment in MobMess.

Other methods have recently been developed to cluster thousands of plasmids^{3,34}, but unlike MobMess, they are not designed to identify plasmid systems or analyze metagenomic data. To compare methods, we ran MobMess on the same set of 9,894 reference plasmids analyzed by Redondo-Salvo et al.³ (Figure S10). In their study, Redondo-Salvo et al. constructed a plasmid similarity network with 79,727 edges. However, these edges span a wide range of similarity levels, where 66.5% of edges represent an alignment that covers <90% of either sequence ($\geq 10\%$ is not aligned) and 19.0% of edges have <70% alignment coverage ($\geq 30\%$ is not aligned). In contrast, MobMess applies a stricter threshold of $\geq 90\%$ coverage to construct a smaller but more refined

set of 39,680 edges (connecting 25,270 unique pairs of plasmids). Moreover, Redondo-Salvo et al.'s edges are undirected, while MobMess's edges are directed to track smaller versus larger sequences. Retaining this extra information allowed MobMess to distinguish between the 10,860 pairs (43.0%) with unidirectional connections, representing a backbone contained in a compound plasmid, versus the 14,410 pairs (57.0%) with bidirectional connections, representing nearly identical plasmids.

Besides network construction, these methods also diverge in how they conceptually organize plasmids. MobMess dereplicates the 9,894 plasmids into 7,132 non-redundant sequences and then organizes them into 1,044 plasmid systems. In contrast, Redondo-Salvo et al. identified 641 clusters, or 'PTUs'³. We found that 135 PTUs did correspond one-to-one to a plasmid system in MobMess, but the other PTUs spanned a wide range of evolutionary relations. At one extreme, 251 PTUs were simple sets of nearly identical plasmids, representing recent and strong relations. At the other extreme, 45 PTUs were complex mixtures of distinct plasmid systems, representing distant and weak relations. For example, the largest PTU contained 2,460 plasmids, which MobMess further dissected into 1,481 non-redundant plasmids and 461 plasmid systems. Figure S11 demonstrates one such plasmid system, where MobMess precisely connects the system's backbone to its compound plasmids in a "star"-like topology, while the approach by Redondo-Salvo et al. connects almost every pair of these plasmids to each other, which obfuscates the internal organization of the plasmid system. Perhaps this is in part because the method by Redondo-Salvo et al. and another related method by Acman et al.³⁴ have only been tested on reference plasmids that have been completely assembled, while MobMess is designed to handle metagenomic data by distinguishing between fragmented versus complete (circular) plasmids.

Supplementary References

1. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
2. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
3. Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P. & de la Cruz, F. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
4. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
5. Heer, J., Card, S. K. & Landay, J. A. prefuse: a toolkit for interactive information visualization. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 421–430 (Association for Computing Machinery, 2005).
6. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C. & Wang, J. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 1–13 (2015).
7. David, L. A., Weil, A., Ryan, E. T., Calderwood, S. B., Harris, J. B., Chowdhury, F., Begum, Y., Qadri, F., LaRocque, R. C. & Turnbaugh, P. J. Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *MBio* **6**, (2015).
8. Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., Frenette, J., Boudreau, D. K., Simard, J.-L., Chabot, I., Domingo, M.-C., Trottier, S., Boissinot, M., Huletsky, A., Roy, P. H., Ouellette, M.,

- Bergeron, M. G. & Corbeil, J. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10**, 707 (2016).
9. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D. & Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
 10. Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., Li, H., Wu, C., Hu, C., Xu, Q., Zhou, J., Cai, S., Wang, D., Huang, Y., Breban, M., Qin, N. & Ehrlich, S. D. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18**, (2017).
 11. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J. M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H. B., Juncker, A. S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E. G., Brunak, S., Clément, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W. M., Zucker, J. D., Raes, J., Hansen, T., Bork, P., Wang, J., Ehrlich, S. D. & Pedersen, O. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, (2013).
 12. Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K. J., Jackson, M. A., Xia, Y., Chen, X., Chen, B., Xia, H., Xu, C., Li, F., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Wang, J., Steves, C. J., Bell, J. T., Li, J., Spector, T. D. & Jia, H. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut

Microbiome. *Cell systems* **3**, 572 (2016).

13. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C. & Segata, N. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
14. Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., Birren, B. W., Xavier, R. J., Blainey, P. C., Singh, A. K., Gevers, D. & Alm, E. J. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435 (2016).
15. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E. & Segal, E. Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
16. Rampelli, S., Schnorr, S. L., Consolandi, C., Turrioni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A. N., Henry, A. G. & Candela, M. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
17. Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xiaoxia, X. I., Liang, Z., Hou, Q., Zhou, B., Qin, N. & Zhang, H. Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, (2016).
18. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J. Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H. B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Doré, J., Ehrlich, S. D., Bork, P. & Wang, J. An integrated

- catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
19. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
 20. Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Xu, Z. Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarreal, O., Foster, M., Gujja-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T. & Lewis, C. M. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 1–9 (2015).
 21. Fallon, A. M., Kurtz, C. M. & Carroll, E. M. The oxidizing agent, paraquat, is more toxic to *Wolbachia* than to mosquito host cells. *In Vitro Cell. Dev. Biol. Anim.* **49**, 501–507 (2013).
 22. Uribe-Alvarez, C., Chiquete-Félix, N., Morales-García, L., Bohórquez-Hernández, A., Delgado-Buenrostro, N. L., Vaca, L., Peña, A. & Uribe-Carvajal, S. *Wolbachia pipientis* grows in *Saccharomyces cerevisiae* evoking early death of the host and deregulation of mitochondrial metabolism. *Microbiologyopen* **8**, e00675 (2019).
 23. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**, (2018).
 24. Chen, Z., Zhong, L., Shen, M., Fang, P. & Qin, Z. Characterization of *Streptomyces* plasmid-phage pFP4 and its evolutionary implications. *Plasmid* **68**, 170–178 (2012).
 25. Oliva, M. A., Martin-Galiano, A. J., Sakaguchi, Y. & Andreu, J. M. Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).
 26. Dokland, T. Molecular Piracy: Redirection of Bacteriophage Capsid Assembly by Mobile Genetic Elements. *Viruses* **11**, (2019).
 27. Pfeifer, E., Moura de Sousa, J. A., Touchon, M. & Rocha, E. P. C. Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).

28. Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.* **16**, e1007781 (2020).
29. Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z. & Zhu, H. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8**, (2019).
30. Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T. & Goesmann, A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* **6**, (2020).
31. Andreopoulos, W. B., Geller, A. M., Lucke, M., Balewski, J., Clum, A., Ivanova, N. N. & Levy, A. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Res.* **50**, e17 (2022).
32. Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z. & Ou, H.-Y. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* **47**, D660–D665 (2019).
33. Pesesky, M. W., Tilley, R. & Beck, D. A. C. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid* **102**, 10–18 (2019).
34. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 1–11 (2020).
35. Olm, M. R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P. B. & Banfield, J. F. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, (2020).
36. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).