# Discovering plasmids in metagenomes based on genetic architecture

Michael K. Yu[1,*], Emily C. Fogarty[2,3,*], A. Murat Eren[2,3,4]

1. Toyota Technological Institute at Chicago, Chicago, IL 60605, USA
2. Department of Medicine, University of Chicago, Chicago, IL 60637, USA
3. Graduate Program in the Biological Sciences, The University of Chicago, Chicago, IL 60637, USA
4. Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

* co-first authors

## ABSTRACT

Plasmids play a critical role in rapid bacterial adaptation by encoding accessory functions that may increase the host's fitness. However, the diversity and ecology of plasmids is poorly understood due to computational and experimental challenges in plasmid identification. Here, we report the Plasmid Classification System (PCS), a machine learning classifier that recognizes plasmid sequences based on gene functions. To train PCS, we performed a large-scale discovery and comparison of gene functions in a reference set of >16,000 plasmids and >14,000 chromosomes. PCS accurately recognizes a diverse range of plasmid subtypes, and it outperforms the previous state-of-the-art approach based on k-mer decomposition of sequences. Armed with this model, we conducted, to our knowledge, the largest search for naturally occurring human gut plasmids in 406 publicly available metagenomes representing 5 countries. This search yielded 6,257 high-confidence predicted plasmids, of which 576 had evidence of a circular conformation based on pair-end mapping. These predicted plasmids were found to be highly prevalent across the metagenomes compared to the reference set of known plasmids, suggesting there is extensive and uncharacterized plasmid diversity in the human gut microbiome.

## INTRODUCTION

The human body is home to an astonishing number of microbes, collectively encoding ten times more genes than the human genome itself (Ursell et al. 2014). Through this extensive genetic diversity, gut microbes play critical roles in our well-being by extracting energy from dietary nutrients (Schwalm and Groisman 2017), synthesizing vitamins (Goodman et al. 2009), offering protection against pathogens (Zeng et al. 2016), and promoting immune homeostasis (Peterson et al. 2015). Changes in these microbial communities have been associated with a multitude of diseases, including inflammatory bowel disease (Vila et al. 2018), metabolic disorders (Bäckhed et al. 2004), and cancer (Kostic et al. 2013). These associations have largely focused on the presence or absence of microbial taxa in the human gut. However, most microbially-mediated human diseases require much deeper insights into the molecular and functional properties of the gut microbial community that go beyond the level of individual taxa.

A major driver of microbial evolution are plasmids: self-replicating, extrachromosomal DNA that can be exchanged between different bacterial cells. Plasmids can inhabit specific bacterial hosts, or may have a broad host range (A. Jain and Srivastava 2013; Klümper et al. 2015). The acquisition of a plasmid through conjugation or transformation may allow the host to rapidly adapt to changing environmental conditions by expressing the genes that are present on the plasmid (Sentchilo et al. 2013). Plasmids have typically been studied in pathogenic bacteria for their ability to alter virulence (Timothy J. Johnson 2009) and antibiotic resistance (Millan 2018). For example, Enteroinvasive *E. coli* (EIEC) possess a large plasmid called pINV, which encodes the genes necessary for EIEC invasion of human macrophages and epithelial cells (Lan, Stevenson, and Reeves 2003).

However, plasmids also exist in many non-pathogenic organisms that inhabit the human gut. Besides genes involved in pathogenesis, plasmids can carry genes for increased salt tolerance (Broaders et al. 2016), inter-bacterial competition (Millette et al. 2008), and increased metabolic potential (Chassy, Gibson, and Guiffrida 1978; Kankainen et al. 2009). A plasmid can change the fitness of its bacterial host, which may in turn reshape key ecological properties of the gut microbiome. Despite their significant role in bacterial lifestyles, plasmids remain difficult to study, particularly when the fitness advantages gained by the plasmid are not linked to a clear phenotype such as virulence.

Monoculture of bacteria in nutrient-rich media is one approach to studying their genetic content, including plasmids. However, it is challenging to cultivate many gut bacteria, and plasmids are frequently lost from cells when bacteria are grown in laboratory media. To address this bottleneck, (Jones and Marchesi 2006) developed TRACA, a system to physically capture plasmids from environmental samples without the need for culturing. The captured plasmids are replicated in laboratory *E. coli* and can be further characterized in this experimental system. TRACA is effective for plasmid capture, yet is laborious and requires the researcher to have access to the biological sample. Recently, high-throughput sequencing technologies have enabled unprecedented characterization of gut microbiomes through ‘metagenomics’, a strategy that sequences the entire DNA content of a sample (Handelsman 2004). Although metagenomes contain an enormous amount of information, avoid the requirement for culturing, and are often publicly available, it remains a challenge to categorize the resulting DNA sequences according to their origin.

Several computational strategies have been developed to help distinguish between plasmid and chromosomal DNA. One strategy uses machine learning (Zhou and Xu 2010; Krawczyk, Lipinski, and Dziembowski 2018; Pellow, Mizrahi, and Shamir 2020) to learn patterns based on short signatures of typically ~3-7 nucleotides, called k-mers. Another strategy is to identify plasmids based on their circularity during assembly (Antipov et al. 2019; Rozov et al. 2017). Another strategy has been to annotate sequences with known genes related to plasmid replication and conjugative transfer. Plasmidfinder (Carattoli et al. 2014) identifies genes related to plasmid replication, but it is trained on a limited number of plasmids of the family *Enterobactericiae*. A more recent tool, MOB-suite (Robertson and Nash 2018), applies a similar approach using a broader set of plasmid replication and mobilisation genes. Beyond these canonical functions, the full repertoire of genes that distinguishes plasmids from other sequences has not been extensively modeled in a systematic and data-driven manner.

Here, we develop a machine learning approach, the Plasmid Classification System (PCS), to identify plasmids based on all molecular functions encoded in a sequence. We apply PCS on assemblies from 406 human gut metagenomes, revealing a diverse set of new plasmids.

## RESULTS AND DISCUSSION

### Establishing a vocabulary of plasmid and chromosome-enriched gene functions

As a first step in training a plasmid classifier, we ran a large-scale analysis of >50 million genes across a reference set of 16,827 plasmids and 14,367 chromosomal sequences (**Figure 1A**, **Methods**). To understand the functions of these genes, we annotated them to evolutionarily conserved gene families using two different approaches. First, to take advantage of prior biological knowledge, we searched genes for sequence homology with known gene families in two databases: the Cluster of Orthologous Groups of proteins (COG) (Galperin et al. 2015) and Pfam (El-Gebali et al. 2019). Second, to capture uncharacterized functions potentially missed by these databases, we also clustered genes into >1 million *de novo* families using a highly parallelized sequence alignment and clustering tool called mmseqs (Steinegger and Söding 2017). We allowed genes to be assigned to multiple families.

We found that *de novo* families were critical for a comprehensive study of plasmid-enriched functions. Known gene families are biased, explaining functions for only 71% of all plasmid genes plasmids as opposed to 89% of chromosomal genes (**Figure 1B**). Moreover, many plasmids have only a small subset or none of their genes annotated (**Figure 1C**). Incorporating *de novo* families enabled us to characterize 95% of all plasmid genes and a large fraction of genes in any one plasmid. To further understand if these families would be informative for creating a plasmid classifier, we calculated their enrichment in plasmids versus chromosomes. While some of the known families were enriched in plasmids, there were thousands of more *de novo* families that were not only enriched at all but also enriched at stronger levels (**Figure 1D-E**).

### Addressing length and taxonomic biases in reference sequences

We further preprocessed the reference sequences to address three major challenges in discovering new plasmids in metagenomes. The first challenge is that plasmids (~1-100kb, with the exception of megaplasmids) are typically much shorter than chromosomes (~500kb-5Mb), so sequence length would be a highly accurate but not insightful predictor. Second, assembly of metagenomes from short-read sequencing typically results in short contigs that are fragments of the original genome. To address these two challenges, we sliced plasmids and chromosomes into subsequences using a 10kb window with 5kb increments.

The third challenge is that reference plasmids are highly redundant and not a uniform representation of possible genetic diversity. The taxonomic assignments of plasmids are heavily skewed, with 5,974 (35%) from the family *Enterobacteriaceae*. The three most common species assignments are well-studied human pathogens: *E. coli* (2589), *K. pneumoniae* (1465), and *S. enterica* (817). Because taxonomy is missing or possibly incorrect for some plasmids, we also categorized plasmids into XYZ subtypes based on sequence similarity (see **Methods**). This analysis recapitulated a similar skew, with the largest subtype containing 9,524 (57%) of the plasmids on one extreme and 3,971 (19%) plasmids alone in their own subtypes on the other extreme. To address this skew, we assigned fractional weights to the sliced sequences such that the total weight for every subtype was equal.

**PCS: a state-of-the-art plasmid classifier system based on gene functions**

Using 10kb sliced sequences and subtype-based weights, we trained a logistic regression that uses the known and *de novo* families as features to distinguish plasmids from chromosomes. The resulting model, called the Plasmid Classification System (PCS), takes in as input the set of families encoded in a sequence, and it returns a score between 0 to 1 representing the probability of being a plasmid.

We compared PCS versus PlasClass (Pellow, Mizrahi, and Shamir 2020), a recent method that also fits a logistic regression but uses k-mers of length 3-7 as features, in their ability to classify the 10kb sequences. We did not compare with other k-mer methods (Krawczyk, Lipinski, and Dziembowski 2018; Zhou and Xu 2010) because the PlasClass study reported better performance than them. We first evaluated performance in 4-fold cross-validation using a “naive” random splitting of sequences (**Figure 1A**) and a uniform weighting of 10kb slices to calculate precision and recall. PCS achieved a moderately higher area under the precision-call-curve (AUC=0.55) than PlasClass (AUC=0.45, **Figure 2B**). While this type of naive splitting has often been used in microbial classification tasks, it is ill-designed because the existence of sequence subtypes causes the training and test data to have similar sequences. As a more accurate benchmark, we also evaluated using an “informed” split, which keeps sequences from the same subtype together in training or test, and calculated precision and recall using the subtype-based weights. This scenario reveals a greater performance divide between PCS (AUC=0.70) versus PlasClass (AUC=0.17), demonstrating the importance of using gene functions to identify new types of plasmids.

**PCS-predicted plasmids are more prevalent in healthy human gut microbiomes than previously established plasmids.**

Plasmids are important drivers of microbial evolution, and allow for rapid adaptation of their microbial host to new or changing environments. Even in a healthy human gut, a microbe may encounter situations where maintenance and expression of a plasmid is beneficial, for example in dealing with bile salt stressors (Broaders et al. 2016). A plasmid that is conserved across many human guts may provide particularly advantageous traits, or may be able to replicate in a broad range of microbial hosts. We are interested in identifying plasmids that are prevalent across healthy individuals as they may contribute to maintaining a homeostatic environment. Many reference plasmids are human-associated, however, these plasmids are often isolated from individuals with diseases, and may not represent plasmids found in healthy human guts. Here, we examined how our current understanding of plasmids (the reference set of plasmids) compares to the plasmids predicted by PCS with respect to their distribution and relevance across human populations.

We downloaded 406 publicly available human gut metagenomes from Fiji, Tanzania, Italy, China and the USA. We individually assembled each metagenome into contigs using IDBA_UD (Peng et al. 2012). We ran PCS on the 7,787,977 assembled contigs and applied a score cutoff of ≥0.5 to identify plasmids. We depreplicated the predicted plasmids by assigning plasmids with greater than 90% alignment coverage and sequence identity to the same group, and used the longest contig as a representative (see **Methods**), resulting in a set of 31,399 putative plasmids. A similar de-replication was applied on the reference plasmids. We separately recruited reads for all 406 metagenomes to the remaining 11,059 reference plasmids and 31,399 predicted plasmids using bowtie2 (Langmead and Salzberg 2012) and the snakemake (Köster and Rahmann 2012) workflows in anvi'o (Shaiber and Murat Eren 2018). For each plasmid, we computed its detection (the fraction of the plasmid covered by at least one read) across all metagenomes.

To ensure that the contigs were likely to be plasmids, we filtered out all predicted plasmids that had a score of less than 0.90, leaving 6,257 high-confidence plasmids. For both known and predicted plasmids, we only kept those that had a detection greater than 0.95 in at least one of the 406 metagenomes, to ensure all of the plasmids could be considered human gut-associated for fair comparison. Of the 11,059 reference plasmids, only 148 passed this filter, compared to 5,311 of the 6,257 predicted plasmids. However, it was still possible that the 148 reference plasmids would be more prevalent across the 406 metagenomes than the 148 most prevalent predicted plasmids. To identify the distribution of the known and predicted plasmids, we plotted the detection of both sets of 148 plasmids across all metagenomes (**Figure 3**). The detection (between 0 and 1) of each plasmid in each metagenome is shown, and known and PCS-predicted plasmids are individually clustered based on their detection across all samples. 53% (78/148) of the subset of PACS-predicted plasmids is present in greater than 10% of all metagenomes, compared to 7% (11/148) of the reference plasmids.

Within the predicted plasmids there is a clear distinction between plasmids that are prevalent in industrialized versus non-industrialized countries. Italy and Tanzania have few plasmids that are present across most samples, however, these results may be influenced by their lower depth of sequencing. A comparison between the USA and Fiji indicates that there is a geographical stratification of plasmids, similar to what has previously been observed about the geographical distribution of mobile genetic elements (Brito et al. 2016).

These results indicate that if we are interested in studying how plasmids impact microbial evolution in a healthy human gut, we are missing the vast majority of highly conserved plasmids by traditional methods. Using our model, we are able to identify novel plasmids that are maintained across many individuals. Maintenance of a gene or organism across similar environments often indicates that it plays an important role in ecosystem functioning or maintenance of a stable community. This method of identifying plasmids that are conserved across many individuals will allow us to focus on relevant targets for future experimental investigations examining the impacts of plasmids on microbial fitness.

### PCS predicts a wide variety of plasmids

To ensure that PCS was not biased to predicting plasmids of a particular length or genetic content, we manually examined a large subset of the predicted plasmid sequences. Reassuringly, we are able to predict many different types of plasmids. We focused on the analysis of gene content and coverage of four specific plasmids. These plasmids are depicted in Figure 4, with their predicted ORFs and corresponding annotations, and the short read coverage of that plasmid in the metagenome it was assembled from. Consistent read coverage indicates that these plasmids exist in this conformation in the sample, and are not assembly artifacts that are non-specifically recruiting reads. Although we can also predict fragmented plasmids, all four plasmids are predicted to be circular (see **Methods**), indicating that they are fully assembled contigs. The only example of a substantial variation in read coverage can be seen in Plasmid 3, where a transposase recruits twice as many reads as the rest of the plasmid, indicating that it is present in another context in another place in the genome, as could be expected for a transposon.

PCS can predict plasmids with canonical genes, for example Plasmid 1, which contains a replication gene, as well as a toxin-antitoxin system. Toxin-antitoxin systems produce a long lasting toxin and a short lasting anti-toxin, so that cells that lose the plasmid will be killed by the toxin. Toxin-antitoxin systems that force the maintenance of their plasmid are common in many plasmids.

PCS is also able to predict short plasmids (Plasmid 2) that contain only a predicted replication gene. PCS is able to predict larger plasmids that carry many more genes with known and unknown functions (Plasmid 3). The longest plasmid predicted by PCS was 205,445 base pairs.

An interesting set of plasmids predicted by PCS are those that are circular (see **Methods**), indicating a complete assembly, yet do not carry replication genes matching any COG or replication gene in NCBI databases (Plasmid 4). These plasmids may carry novel types of replication genes, or replicate using a different strategy than reference plasmids. Alternatively, these contigs may represent mobile genetic elements with lifestyles that blur the distinctions between categories of horizontally transferred elements. Serendipitous discoveries have given us examples of phages carrying plasmid segregation proteins (Oliva et al. 2012), plasmids encoding capsids ("Characterization of Streptomyces Plasmid-Phage pFP4 and Its Evolutionary Implications" 2012), and phagemids, which can integrate into the genome like phages or replicate in the cytoplasm like a plasmid (Dokland 2019). A targeted effort to understand the spectrum of plasmids to phages may reveal many undiscovered mobile genetic elements that do not fall into defined categories. The insights into novel avenues of plasmid biology generated by PCS will enable experimental biologists to systematically explore new plasmid lifestyles.

## METHODS

### Analysis of reference plasmids and chromosomes

We obtained a reference list of NCBI accessions of plasmids from the March 5, 2019 version of PLSDB (Galata et al. 2019). A collection of complete bacterial genomes were download on October 26, 2019 from NCBI RefSeq database using the instructions at https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete. In the bacterial genomes, we added the replicons that were labeled as plasmids in the metadata to the set of plasmids from PLSDB, resulting in a total set of 16,827 plasmids. Other replicons in the bacterial genomes formed the set of 14,637 chromosomal sequences.

To infer plasmid and chromosome subtypes, we ran `mash dist` (sketch size 100000, kmer size 21) to calculate a distance score of 0 to 1 between every pair of sequences (Ondov et al. 2016). Conceptually, these distances form an undirected graph. To identify highly similar sequences, we applied a threshold of 0.1 on the distances to produce an undirected graph connecting sequences. We defined a “subtype” as one of the 7,326 connected components in the graph. As some of these components contain both plasmids and chromosomes, there were 3,971 subtypes containing at least plasmid, and 3,391 subtypes containing at least one chromosome.

We annotated genes in the reference plasmids and chromosomes using the “contigs” snakemake (Köster and Rahmann 2012) workflow in anvi’o (Shaiber and Murat Eren 2018; Eren et al. 2015). This workflow first identified >50 million protein-coding genes and their corresponding amino acid sequence using Prodigal (Hyatt et al. 2010). Next, it searched these sequences for homology in Cluster of Orthologous Groups of proteins (COG) (Galperin et al. 2015) using diamond (Buchfink, Xie, and Huson 2015), an accelerated blast-like tool. It also searched for homology with profile HMMs in the Protein Family Database (Pfam) (El-Gebali et al. 2019) using the hmmscan command in the HMMER software package (v3.3, hmmer.org). For both searches, the default e-value cutoff of $10^{-10}$ by anvi’o was used.

Additionally, we inferred a set of *de novo* gene families by applying mmseqs version 10.6d92c (Steinegger and Söding 2018) on all genes. As a pre-processing step, we first ran the `mmseqs clusthash` command to collapse identical amino acid sequences into a non-redundant set for faster downstream analysis; the collapsing was inverted afterwards to annotate all genes. Next, we ran the `mmseqs cluster`, which computes all pairwise alignments above a minimum sequence identity threshold and then clusters genes using a greedy linear-time algorithm. We set the identity threshold `--min-seq-id` to either 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, or 0.05 to infer a wide range of possible families. Afterwards, we merged nested families, i.e. if family X contains all the genes in family Y, then we only keep X. We also discarded any family for which all of its genes are in only one sequence. The final result was 1,090,132 families with 162,783,114 annotations to the genes (genes may have multiple annotations). This analysis took advantage of mmseqs built-in parallelism, taking ~6 hours using 256 CPU cores.

### Training and evaluation of PCS

We trained a logistic regression with elastic net regularization using the LogisticRegression class from the scikit-learn Python package (Pedregosa et al. 2011). We performed a grid search of hyperparameters, with *alpha* ranging from $10^{-8}$ and $10^{-3}$ in multiplicative increments of np.log10(2) and *l1_ratio* being 0.0, 0.25, 0.5, 0.75, or 1.0. The best hyperparameters were chosen separately for the two scenarios in **Figure 3B-C**. For the "informed" split and sample weighting scenario (**Figure 3C**), the best values were *alpha*=3.16x10$^{-6}$ and *l1_ratio*=0.0. We used this setup and values to retrain PCS on all 10kb sliced sequences.

We defined a set of weights on 10kb sliced sequenced to satisfy the following conditions: (a) the total weight of a plasmid (or chromosome) across its 10kb slices is equal to the weight of any other plasmid (or chromosome) in the same subtype, (b) the total weight of a plasmid subtype is the same as any other subtype, (c) the total weight of all plasmid slices equals that of all chromosomal slices, and (d) the total weight across all slices is equal to the number of slices. These conditions result in a unique assignment of weight values.

### Metagenomic datasets, assembly, read recruitment, and profiling

We downloaded the metagenomic datasets from the National Center for Biotechnology Information (NCBI) using the software `fastq-dump`. The countries represented are Tanzania (Rampelli et al. 2015), Italy (Rampelli et al. 2015), China (Qin et al. 2012), Fiji (Brito et al. 2016) and the USA (Turnbaugh et al. 2007; Obregon-Tito et al. 2015).

All steps of quality filtering, metagenomic assembly, read recruitment and profiling were automated using snakemake (Köster and Rahmann 2012) workflows in anvi'o (Shaiber and Murat Eren 2018). We used the `ilumina-utils` (Murat Eren et al. 2013) commands `iu-gen-configs` and `iu-filter-quality-minoche` with the flag `--ignore-deflines` to quality filter the raw paired-end reads. We assembled each metagenome individually using IDBA_UD (Peng et al. 2012) with default settings except the flag `--min_contig 1000`. To calculate the coverage of each contig within its respective metagenome, we recruited the reads from that metagenome back to the assembled contigs using bowtie2 v.2.0.5 (Langmead and Salzberg 2012) with default parameters. We converted the SAM files into BAM files using samtools v1.3.1 (Li et al. 2009). We generated profile databases from the BAM files using anvi'o (Eren et al. 2015), which stores coverage information and allows for visualization of the coverage plots. The per nucleotide coverage values were exported from the profile database with the command `anvi-get-split-coverages`. The coverage values were averaged every 20 base pairs and used to create circular coverage plots in anvi'o (Eren et al. 2015).

To identify circular contigs, we examined the reads that mapped onto the predicted plasmids in a paired-end configuration. While paired ends typically map near each other in a contig, some should map in a "reverse-forward" (RF) configuration at opposite ends of a circular contig as an artifact of how the contig was linearized. We deemed a contig as circular if (a) there were at least 5 RF paired ends with a distance (insert length) ≥500 and (b) the median insert length of such paired ends satisfying (a) is ≥80% of the contig's length.

### Predicting plasmids from metagenomic assemblies

We followed the same procedure as the reference plasmids and chromosomes to annotate known gene families in all contigs in the metagenomic assemblies. To annotate *de novo* families, we first converted every *de novo* family into a position-specific scoring matrix, which captures the observed sequence variation of genes in this family, using `mmseqs result2profile` (default parameters). We then used `mmseqs search` (default parameters) to search for genes across the assembly contigs against the profile. We kept hits for which the alignment coverage was ≥80% in both the assembly gene and the profile and where the identity was at least ≥X-0.05 where X is the minimum identity threshold used to infer the family. For example, if a family was inferred using a threshold of 0.8, then we kept hits with an identity ≥0.75. Finally, we ran the trained plasmid classifier based on these known and *de novo* annotations to assign a score to every contig.

### Determining the detection of the known and predicted plasmids across metagenomes

PCS assigns a plasmid prediction score between 0 and 1 to each contig. We dereplicated these sequences by calculating the average nucleotide identity (ANI) between every pair of predicted plasmids using FastANI (--fragLen 1000, -k 16) (C. Jain et al. 2018). We defined groups of highly similar plasmids as the connected components in an undirected graph consisting of pairs (s, t), for which the FastANI alignment had an ANI≥90% and covered ≥90% of both s and t. We took the longest plasmid in every cluster, resulting in 31,399 de-replicated plasmids. We used the same approach to dereplicate the reference plasmids. To ensure we were using the contigs most likely to be plasmids, we filtered out the predicted plasmids with a score of less than 0.90, resulting in 6,257 remaining plasmid sequences. We recruited reads from all 406 metagenomes to each plasmid set (reference and predicted) separately, using the same software and approaches as we had for recruiting reads to the assembled contigs. Detection of each contig is a value between 0 and 1 that indicates the proportion of the contig that had at least one read map to it. Detection is automatically calculated during profiling in anvi'o. We accessed the detection data using `anvi-export-table`, and used these tables to generate the heatmaps in **Figure 3**.

### Plasmid clustering by detection

We used the built in functionality of heatmap.2 to cluster the plasmids based on their detection across all metagenomes. The dendrograms were generated using Euclidean distance and median linkage.

### Manual plasmid gene annotation

We ran the program `anvi-gen-contigs-database` and `anvi-run-ncbi-cogs` on the fasta files containing the dereplicated known and predicted plasmids. `anvi-gen-contigs-database` runs Prodigal v2.6.3 (Hyatt et al. 2010), which identifies open reading frames (ORFs) in contigs. `anvi-run-ncbi-cogs` identifies COG functions (Galperin et al. 2015) from the ORFs predicted by Prodigal. We manually imported the COG functions from the anvi'o interactive interface into the plasmid maps produced by snapgene (Insightful Science; snapgene.com). Any gene that did not have a COG function assigned was manually curated using NCBI BLASTx.

### Data visualization

We used the R package `heatmap.2` to generate the raw heatmap and dendrogram, and inkscape to refine the figure. We used anvi'o to create the circular coverage plots, and snapgene to create the plasmid maps with labeled genes. These were combined in inkscape 1.0.1.

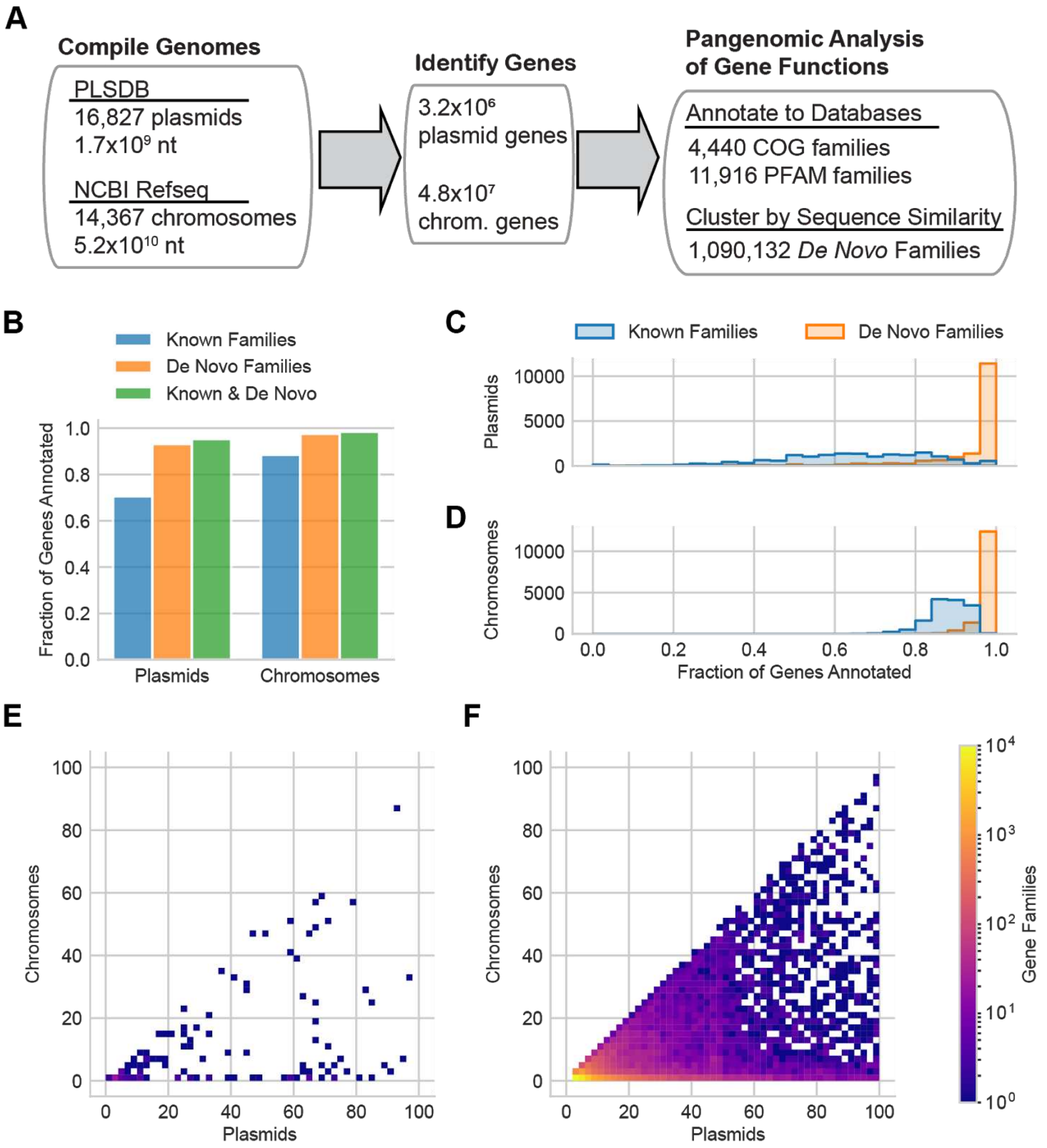

**Figure 1. Establishing a vocabulary of gene functions in plasmids and chromosomes. (A)** A pangenomics workflow was applied to characterize gene functions in a reference set of 16,827 plasmids from the PLSDB database (Galata et al. 2019) and 14,367 complete chromosomes from NCBI RefSeq. Across these sequences, >50 million protein-coding genes were identified using Prodigal (Hyatt et al. 2010). To infer function, every gene was assigned to one or more gene families, each representing an evolutionarily conserved amino acid sequence. A known set of gene families consisted of COGs (Galperin et al. 2015) and Pfam (El-Gebali et

al. 2019). A set of gene families were also inferred *de novo* by pairwise alignment of amino acid sequences followed by clustering using a highly parallelized tool (Steinegger and Söding 2017). **(B)** Fraction of all plasmids or all chromosomal genes that are annotated by using known families (blue), de novo families (orange), or a combination of both (green). **(C-D)** Histograms showing the number of plasmid (C) or chromosomal (D) sequences as a function of genes annotated using known or de novo families. **(E-F)** 2-D histograms showing the number of known (E) or de novo (F) gene families as a function of the number of plasmid and chromosomal sequences that contain a family. The number of gene families is log-scaled.

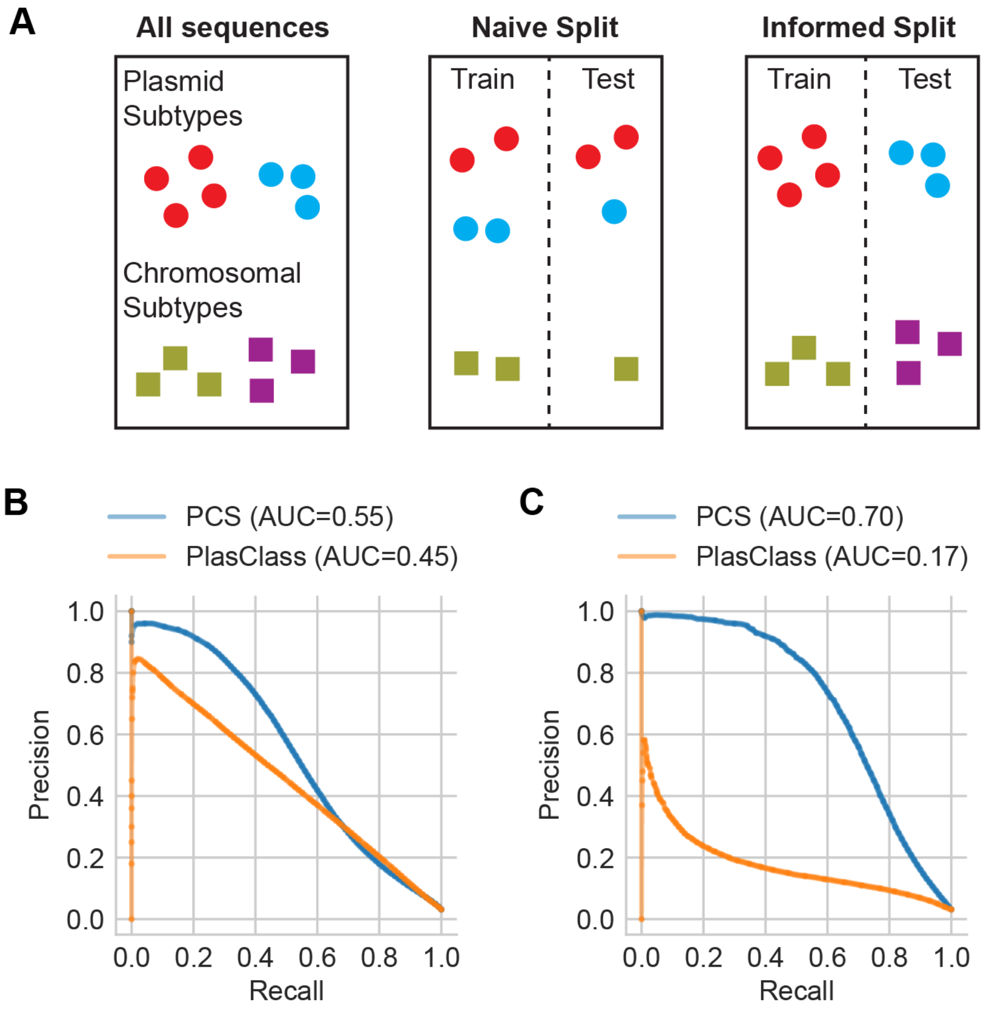

**Figure 2. Evaluation of PCS model performance. (A)** Diagrams of different training-test split configurations. A random "naive" split of plasmids and chromosomal sequences would result in training and test sets that have similar sequences, due to the existence of plasmid and chromosomal subtypes that each 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. **(B-C)** We evaluated the performance of PCS and PlasClass (Pellow, Mizrahi, and Shamir 2020) in 4-fold cross validation. We calculated the precision-recall curve and its area-under-the-curve (AUC) using a naive split and uniform sample weighting **(B)** or an informed split with sample weights that assign plasmid and chromosomal subtypes equal representation **(C)** (see **Methods**).

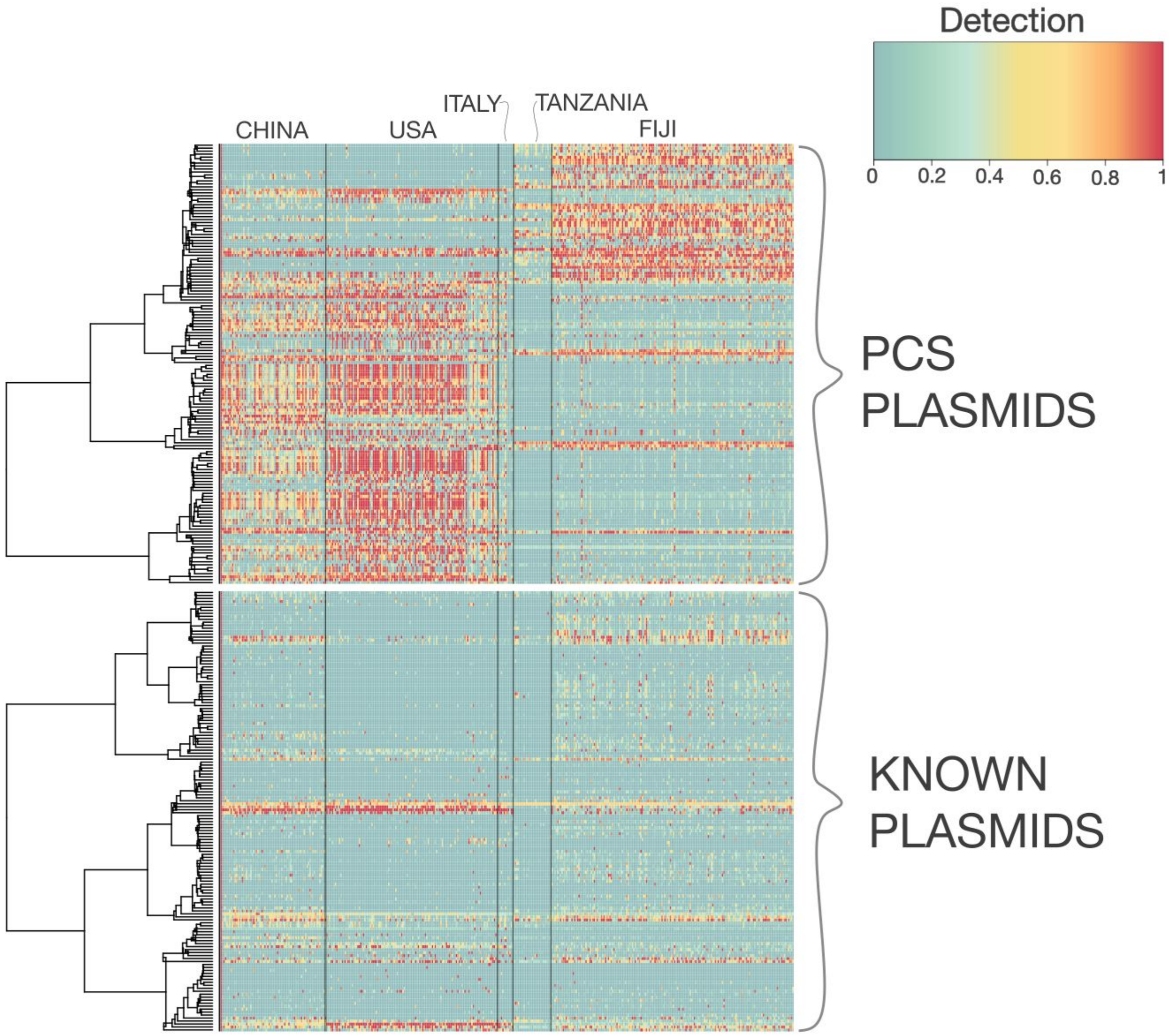

**Figure 3. PCS-predicted plasmids are more prevalent than previously established plasmids across globally distributed human populations.** We recruited reads from 406 globally distributed metagenomes to the collections of known and predicted plasmids. Only plasmids with greater than 0.95 detection in at least one metagenome are shown. Red indicates the plasmid was highly detected, green indicates no detection. Plasmids are hierarchically clustered based on detection values across metagenomes using Euclidean distance and median linkage.

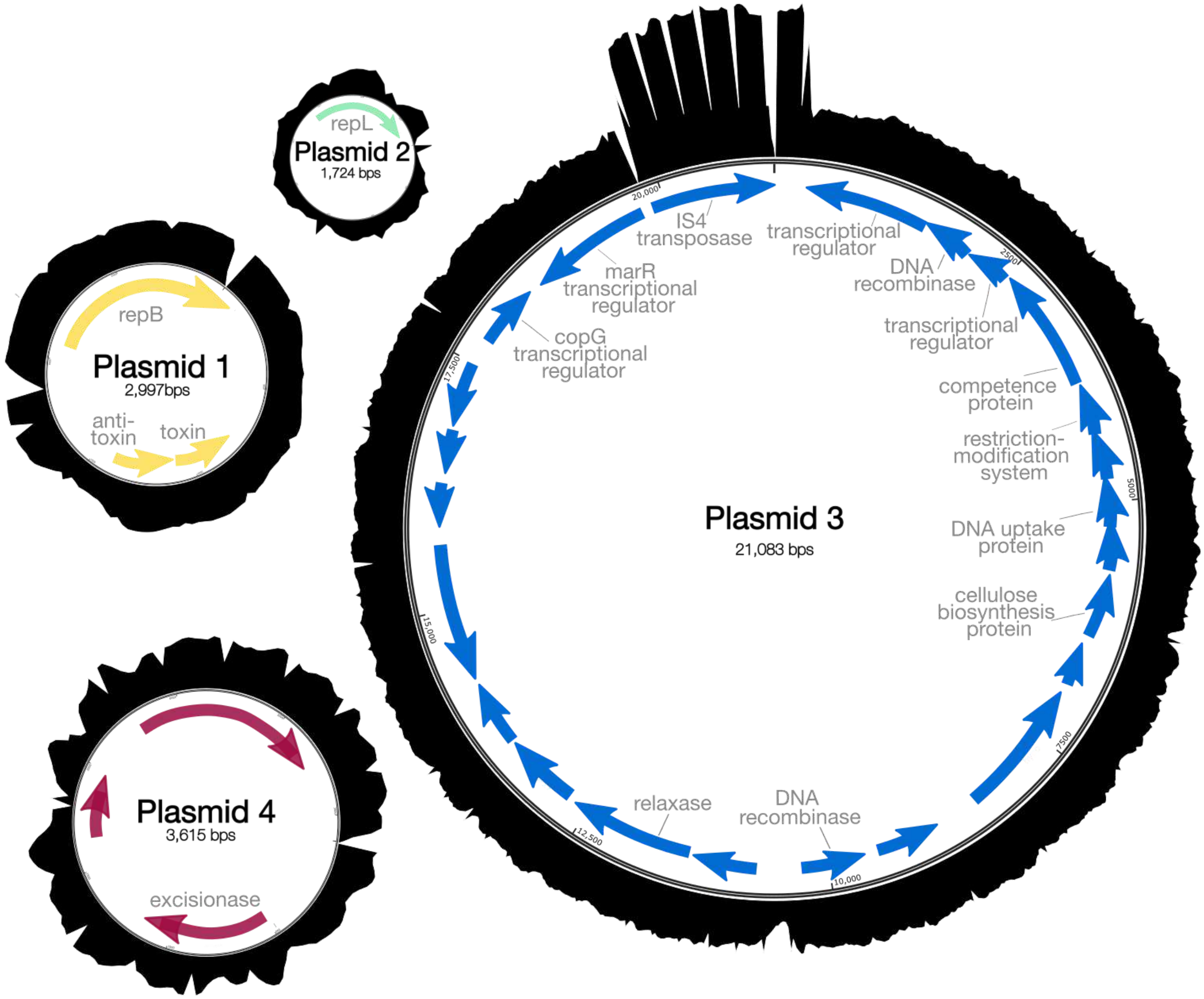

**Figure 4. PCS can predict a variety of plasmids.** Four plasmids with gene annotations and coverage plots based on short read mapping from the metagenome they were initially assembled from. All four plasmids are predicted to be circular. All uncharacterized genes are left unannotated. **Plasmid 1:** plasmid with canonical replication and toxin/anti-toxin genes. **Plasmid 2:** 1.7kb plasmid containing only repL. **Plasmid 3:** 21kb plasmid containing genes of known and unknown function. **Plasmid 4:** non-canonical plasmid that does not contain a discernible replication gene.

## REFERENCES

Antipov, Dmitry, Mikhail Raiko, Alla Lapidus, and Pavel A. Pevzner. 2019. "Plasmid Detection and Assembly in Genomic and Metagenomic Data Sets." *Genome Research* 29 (6): 961–68.

Bäckhed, Fredrik, Hao Ding, Ting Wang, Lora V. Hooper, Gou Young Koh, Andras Nagy, Clay F. Semenkovich, and Jeffrey I. Gordon. 2004. "The Gut Microbiota as an Environmental Factor That Regulates Fat Storage." *Proceedings of the National Academy of Sciences of the United States of America* 101 (44): 15718–23.

Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, et al. 2016. "Mobile Genes in the Human Microbiome Are Structured from Global to Individual Scales." *Nature* 535 (7612): 435.

Broaders, Eileen, Ciarán O'Brien, Cormac G. M. Gahan, and Julian R. Marchesi. 2016. "Evidence for Plasmid-Mediated Salt Tolerance in the Human Gut Microbiome and Potential Mechanisms." *FEMS Microbiology Ecology* 92 (3). https://doi.org/10.1093/femsec/fiw019.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60.

Carattoli, Alessandra, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. 2014. "In SilicoDetection and Typing of Plasmids Using PlasmidFinder and Plasmid Multilocus Sequence Typing." *Antimicrobial Agents and Chemotherapy*. https://doi.org/10.1128/aac.02412-14.

"Characterization of Streptomyces Plasmid-Phage pFP4 and Its Evolutionary Implications." 2012. *Plasmid* 68 (3): 170–78.

Chassy, Bruce M., Evelyn M. Gibson, and Alfred Guiffrida. 1978. "Evidence for Plasmid-Associated Lactose Metabolism in Lactobacillus Casei Subsp. Casei." *Current Microbiology* 1 (3): 141–44.

Dokland, Terje. 2019. "Molecular Piracy: Redirection of Bacteriophage Capsid Assembly by Mobile Genetic Elements." *Viruses* 11 (11). https://doi.org/10.3390/v11111003.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. "The Pfam Protein Families Database in 2019." *Nucleic Acids Research* 47 (D1): D427–32.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data." *PeerJ* 3 (October): e1319.

Galata, Valentina, Tobias Fehlmann, Christina Backes, and Andreas Keller. 2019. "PLSDB: A Resource of Complete Bacterial Plasmids." *Nucleic Acids Research* 47 (D1): D195–202.

Galperin, Michael Y., Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. 2015. "Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database." *Nucleic Acids Research* 43 (Database issue): D261–69.

Goodman, Andrew L., Nathan P. McNulty, Yue Zhao, Douglas Leip, Robi D. Mitra, Catherine A. Lozupone, Rob Knight, and Jeffrey I. Gordon. 2009. "Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat." *Cell Host & Microbe* 6 (3): 279–89.

Handelsman, Jo. 2004. "Metagenomics: Application of Genomics to Uncultured Microorganisms." *Microbiology and Molecular Biology Reviews: MMBR* 68 (4): 669–85.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.

Jain, Aayushi, and Preeti Srivastava. 2013. "Broad Host Range Plasmids." *FEMS Microbiology Letters* 348 (2): 87–96.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9 (1): 5114.

Jones, Brian V., and Julian R. Marchesi. 2006. “Transposon-Aided Capture (TRACA) of Plasmids Resident in the Human Gut Mobile Metagenome.” *Nature Methods* 4 (1): 55–61.

Kankainen, Matti, Lars Paulin, Soile Tynkkynen, Ingemar von Ossowski, Justus Reunanen, Pasi Partanen, Reetta Satokari, et al. 2009. “Comparative Genomic Analysis of Lactobacillus Rhamnosus GG Reveals Pili Containing a Human- Mucus Binding Protein.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (40): 17193–98.

Klümper, U., L. Riber, A. Dechesne, A. Sannazzarro, L. H. Hansen, S. J. Sørensen, and B. F. Smets. 2015. “Broad Host Range Plasmids Can Invade an Unexpectedly Diverse Fraction of a Soil Bacterial Community.” *The ISME Journal* 9 (4). https://doi.org/10.1038/ismej.2014.191.

Köster, Johannes, and Sven Rahmann. 2012. “Snakemake—a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 28 (19): 2520–22.

Kostic, A. D., E. Chun, L. Robertson, J. N. Glickman, C. A. Gallini, M. Michaud, T. E. Clancy, et al. 2013. “Fusobacterium Nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment.” *Cell Host & Microbe* 14 (2). https://doi.org/10.1016/j.chom.2013.07.007.

Krawczyk, Pawel S., Leszek Lipinski, and Andrzej Dziembowski. 2018. “PlasFlow: Predicting Plasmid Sequences in Metagenomic Data Using Genome Signatures.” *Nucleic Acids Research* 46 (6): e35.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357.

Lan, Ruiting, Gordon Stevenson, and Peter R. Reeves. 2003. “Comparison of Two Major Forms of the Shigella Virulence Plasmid pINV: Positive Selection Is a Major Force Driving the Divergence.” *Infection and Immunity* 71 (11): 6298.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16). https://doi.org/10.1093/bioinformatics/btp352.

Millan, Alvaro San. 2018. “Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context.” *Trends in Microbiology* 26 (12): 978–85.

Millette, M., C. Dupont, F. Shareck, M. T. Ruiz, D. Archambault, and M. Lacroix. 2008. “Purification and Identification of the Pediocin Produced by Pediococcus Acidilactici MM33, a New Human Intestinal Strain.” *Journal of Applied Microbiology* 104 (1). https://doi.org/10.1111/j.1365-2672.2007.03583.x.

Murat Eren, A., Joseph H. Vineis, Hilary G. Morrison, and Mitchell L. Sogin. 2013. “A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology.” *PloS One* 8 (6): e66643.

Obregon-Tito, Alexandra J., Raul Y. Tito, Jessica Metcalf, Krithivasan Sankaranarayanan, Jose C. Clemente, Luke K. Ursell, Zhenjiang Zech Xu, et al. 2015. “Subsistence Strategies in Traditional Societies Distinguish Gut Microbiomes.” *Nature Communications* 6 (1): 1–9.

Oliva, María A., Antonio J. Martin-Galiano, Yoshihiko Sakaguchi, and José M. Andreu. 2012. “Tubulin Homolog TubZ in a Phage-Encoded Partition System.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (20): 7711.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 132.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12: 2825–30.

Pellow, David, Itzik Mizrahi, and Ron Shamir. 2020. “PlasClass Improves Plasmid Sequence Classification.” *PLoS Computational Biology* 16 (4): e1007781.

Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics* 28 (11): 1420–28.

Peterson, C. T., V. Sharma, L. Elmén, and S. N. Peterson. 2015. “Immune Homeostasis, Dysbiosis and Therapeutic Modulation of the Gut Microbiota.” *Clinical and Experimental Immunology* 179 (3). https://doi.org/10.1111/cei.12474.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. “A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes.” *Nature* 490 (7418): 55–60.

Rampelli, Simone, Stephanie L. Schnorr, Clarissa Consolandi, Silvia Turroni, Marco Severgnini, Clelia Peano, Patrizia Brigidi, Alyssa N. Crittenden, Amanda G. Henry, and Marco Candela. 2015. “Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.” *Current Biology: CB* 25 (13): 1682–93.

Robertson, James, and John H. E. Nash. 2018. “MOB-Suite: Software Tools for Clustering, Reconstruction and Typing of Plasmids from Draft Assemblies.” *Microbial Genomics* 4 (8). https://doi.org/10.1099/mgen.0.000206.

Rozov, Roye, Aya Brown Kav, David Bogumil, Naama Shterzer, Eran Halperin, Itzhak Mizrahi, and Ron Shamir. 2017. “Recycler: An Algorithm for Detecting Plasmids from de Novo Assembly Graphs.” *Bioinformatics* 33 (4): 475–82.

Schwalm, N. D., and E. A. Groisman. 2017. “Navigating the Gut Buffet: Control of Polysaccharide Utilization in Bacteroides Spp.” *Trends in Microbiology* 25 (12). https://doi.org/10.1016/j.tim.2017.06.009.

Sentchilo, Vladimir, Antonia P. Mayer, Lionel Guy, Ryo Miyazaki, Susannah Green Tringe, Kerrie Barry, Stephanie Malfatti, Alexander Goessmann, Marc Robinson-Rechavi, and Jan R. van der Meer. 2013. “Community-Wide Plasmid Gene Mobilization and Selection.” *The ISME Journal* 7 (6): 1173–86.

Shaiber, Alon, and A. Murat Eren. 2018. “Anvi’o Snakemake Workflows.” July 9, 2018. http://merenlab.org/2018/07/09/anvio-snakemake-workflows/.

Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology*. https://doi.org/10.1038/nbt.3988.

———. 2018. “Clustering Huge Protein Sequence Sets in Linear Time.” *Nature Communications* 9 (1): 1–8.

Timothy J. Johnson, Lisa K. Nolan. 2009. “Pathogenomics of the Virulence Plasmids of Escherichia Coli.” *Microbiology and Molecular Biology Reviews: MMBR* 73 (4): 750.

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. 2007. “The Human Microbiome Project.” *Nature* 449 (7164): 804–10.

Ursell, Luke K., Henry J. Haiser, Will Van Treuren, Neha Garg, Lavanya Reddivari, Jairam Vanamala, Pieter C. Dorrestein, Peter J. Turnbaugh, and Rob Knight. 2014. “The Intestinal Metabolome: An Intersection between Microbiota and Host.” *Gastroenterology* 146 (6): 1470–76.

Vila, Arnau Vich, Floris Imhann, Valerie Collij, Soesma A. Jankipersadsing, Thomas Gurry, Zlatan Mujagic, Alexander Kurilshikov, et al. 2018. “Gut Microbiota Composition and Functional Changes in Inflammatory Bowel Disease and Irritable Bowel Syndrome.” *Science Translational Medicine* 10 (472). https://doi.org/10.1126/scitranslmed.aap8914.

Zeng, M. Y., D. Cisalpino, S. Varadarajan, J. Hellman, H. S. Warren, M. Cascalho, N. Inohara, and G. Núñez. 2016. “Gut Microbiota-Induced Immunoglobulin G Controls Systemic Infection by Symbiotic Bacteria and Pathogens.” *Immunity* 44 (3). https://doi.org/10.1016/j.immuni.2016.02.006.

Zhou, Fengfeng, and Ying Xu. 2010. “cBar: A Computer Program to Distinguish Plasmid-Derived from Chromosome-Derived Sequence Fragments in Metagenomics Data.” *Bioinformatics* 26 (16): 2051–52.