SCISPACE
formerly Typeset

Open access • Posted Content • DOI:10.1101/2021.09.27.462084

PlasmidHostFinder: Prediction of plasmid hosts using random forest — Source link

Derya Aytan-Aktug, Philip T. L. C. Clausen, Judit Szarvas, Patrick Munk ...+7 more authors

Institutions: Technical University of Denmark, University of Chicago, Argonne National Laboratory

Published on: 29 Sep 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Plasmid and Extrachromosomal DNA

Related papers:

- Plasmid Detection, Characterization, and Ecology.
- Evolution of plasmids and evolution of virulence and antibiotic-resistance plasmids.
- Modeling the ecology of parasitic plasmids.
- Plasmids and Their Hosts
- Multi-plasmid clash in a bacterial community: plasmid viability depends on the ecological setting of hosts

Share this paper:

View more about this paper here: https://typeset.io/papers/plasmidhostfinder-prediction-of-plasmid-hosts-using-random-50x6babwo4

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PlasmidHostFinder: Prediction of plasmid hosts using random forest

Derya Aytan-Aktug[1#], Philip TLC Clausen[1], Judit Szarvas[1], Patrick Munk[1], Saria Otani[1], Marcus Nguyen[2,3], James J Davis[2,3,4], Ole Lund[1], Frank M Aarestrup[1].

[1] National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Consortium for Advanced Science and Engineering, University of Chicago, Chicago, Illinois, USA

[3] Data Science and Learning Division, Argonne National Laboratory, Argonne, Illinois, USA

[4] Northwestern Argonne Institute for Science and Engineering, Evanston, Illinois, USA

[#]Corresponding author contact information: daytan@dtu.dk

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ABSTRACT

Plasmids play a major role facilitating the spread of antimicrobial resistance between bacteria. Understanding the host range and dissemination trajectories of plasmids is critical for surveillance and prevention of antimicrobial resistance. Identification of plasmid host ranges could be improved using automated pattern detection methods, compared to homology-based methods due to the diversity and genetic plasticity of plasmids. In this study, we developed a method for predicting the host range of plasmids based on the random forest machine learning method. We trained the models with 8,519 plasmids from 359 different bacterial species per taxonomic level, where the models achieved 0.662 and 0.867 Matthews correlation coefficients at the species and order levels, respectively. Our results suggest that despite the diverse nature and genetic plasticity of plasmids, our random forest model can accurately distinguish between plasmid hosts. This tool can be used online through Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/PlasmidHostFinder/).

Importance:

Antimicrobial resistance is a global health threat to humans and animals causing high mortality and morbidity, and effectively ending decades of success in fighting against bacterial infections. Plasmids confer extra genetic capabilities to the host organisms through accessory genes, which can encode antimicrobial resistance and virulence factors. In addition to lateral inheritance, plasmids can be transferred horizontally between bacterial taxa. Therefore, detecting the host range of plasmids is crucial for understanding and predicting the dissemination trajectories of

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

extrachromosomal genes and bacterial evolution, as well as for taking effective counter measures against antimicrobial resistance.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

INTRODUCTION

Plasmids are extra-chromosomal DNA sequences that have crucial roles in bacterial ecology, evolution and the spread of antimicrobial resistance (AMR) (1). They are typically circular, self-replicating, transferable and tend to obtain, lose or re-arrange their genetic content rapidly which make them extremely mosaic, diverse and plastic. Plasmids are generally composed of backbone and accessory genes. The backbone includes replication (*rep*) and mobility (*mob*) genes which are relatively conserved amongst the plasmids of the same family (2). These features have also been used to type and compare plasmids that are isolated from different hosts using the replicon and MOB typing (3-5). The accessory genes generally confer selective advantages to the host such as AMR, virulence and metal resistance, increasing host survival under stress conditions despite the metabolic costs that plasmids cause to the host (6). Plasmids also harbor toxin-antitoxin systems and act as parasitic entities (7). Plasmids are often competent horizontal gene transfer vectors, and are able to move from one bacterium to another via conjugation, transduction or transformation causing persistent genetic exchange between bacterial hosts (1, 8).

Plasmids vary in the number and range of taxa they can transfer to, replicate in and be maintained in. They can be roughly categorized as having narrow or broad host ranges (9). The features that determine the host range capacity of plasmids are not fully understood yet, but origin of replication, replication initiation dependencies, and origin of transfer are known to be important for host range (9).

Plasmid host ranges can be determined empirically by testing potential hosts *in vitro* (10, 11). However, sequence-based approaches can be used for plasmid host range prediction, which is more

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

practical compared to the empirical methods in terms of turn-around times and usage of laboratory resources (11). Previous studies have attempted to predict plasmid host ranges by comparing oligonucleotide composition of plasmids and chromosomes (1, 9, 10, 12-14). Narrow host range plasmids are expected to have similar oligonucleotide compositions with the host organism due to plasmid sequence amelioration *e.g*. adaptation to a preferred host codon usage (12). However, this method falls short when predicting broad host range plasmids because of plasmids can often transfer to distantly related hosts (9, 12).

Previously developed plasmid identification tools such as PlasmidFinder and PlasFlow have been developed to determine plasmid hosts (3, 15). PlasmidFinder identifies plasmids in whole genome sequences by searching against plasmid replicon sequences from the *Enterobacteriaceae* and Gram-positive species. This alignment-based tool identifies plasmids from these taxa with high accuracy by indicating a source organism based on the best matching replicon. PlasFlow was developed using deep neural network and trained by the *k*-mer counts of fragments at least 1,000 nucleotides in length, and it can detect plasmid hosts at the phylum level. To our knowledge PlasFlow tool is not currently maintained. Recently, Redondo-Salvo et al. (16) developed an automated plasmid classification tool by assigning plasmid taxonomic units (PTUs) using total average nucleotide identity.

Machine learning, a form of artificial intelligence, has been utilized in recent years to understand various biological systems by detecting the linear and non-linear correlations between input and output data (17). It has been used to predict phenotypes and structures in nature, and it has the potential to discover unknown features such as novel AMR genes (18-20). In this study, in order to

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

better predict plasmid hosts and infer plasmid host ranges, we developed a set of random forest-based machine learning models for predicting plasmid hosts at several bacterial taxonomic levels.

MATERIAL AND METHODS

Data set

We downloaded all of the available (10,863) plasmids and corresponding metadata from the Pathosystems Resource Integration Center (PATRIC) (21) in September of 2020. Metadata for each plasmid included the origin of the plasmid and other relevant information such as different database accession numbers, collection date and place, genomic length and features. Four plasmids did not have host information and were removed. The remaining plasmid host information was reported from genus to strain level by the PATRIC database. In total, 1,662 different genus and species level hosts were detected in the plasmid metadata. When fewer than five plasmids had a given host at a given taxonomic level, they were removed. In total, 1,296 under-represented hosts and corresponding plasmids were removed from the dataset to improve the robustness of the models. From the remaining 366 hosts, seven were removed for lacking species annotation. Therefore, we generated machine learning models with 8,519 plasmids and 359 corresponding hosts with species level taxonomy information. The species-level plasmid hosts were assigned to the higher taxonomy levels such as genus, family and order using the NCBI Taxonomy information by the Python ete2 package (version:2.3.10) (22, 23).

Distance tree

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The diversity of the plasmids was measured using an oligonucleotide *k*-mer-based distance tree. The plasmid sequences were indexed using the KMA tool (version: 1.3.9) (24) with the following parameters: -NI -Sparse TG. Next, the *16*-mer Hamming distances were calculated. The distance tree was generated using the CCPhylo tool (version: 0.2.2) using the Neighbor-Joining method (25). The distance tree was visualized using iTOL (version: 4) (26).

K-mer counts

The plasmid genomes were sub-sampled using overlapping *k*-length nucleotides (*k*-mers) and counting the occurrence of every sub-sequence. *K*-mer counting is a well-studied method for analyzing sequence data (27). The sub-sequence size *k* is a critical parameter as the sub-sequences yield various information depending on the size. While short *k*-mers provide information regarding the sequence content, long *k*-mers are informative in detection of unique sequence patterns. We analyzed plasmid genomes using three different *k*-mer sizes: 5, 8 and 10 nucleotides. Counts were calculated using KMC (version: 3.0.0) (28) with the following parameters -fm, -ci1, -cs1677215. These parameters inform the tool regarding the input data format and the minimum and maximum thresholds for the *k*-mer occurrences.

Detection of duplicates

To eliminate possible duplicates from the plasmid collections, we compared the *8*-mer counts of the plasmids to each other. Plasmid pairs with identical *8*-mer counts were treated as duplicates and merged in the dataset. When the pairs had differing host information, the additional hosts were incorporated in the metadata.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sequence length, GC content and codon usage calculation

To capture the plasmid genome characteristics, we calculated the total length of the sequence, GC content and codon usage. Sequence length was calculated by taking all nucleotides into consideration, including ambiguous bases. GC content was calculated by taking the ratio of the total number of cytosine and guanine nucleotides to all nucleotides. Codon usage was determined as the relative frequencies of codons in a coding region which was detected using Prodigal (version: 2.6.3) with default parameters (29).

Model generation and cross validation

For each *k*-mer size, a matrix was generated from *k*-mer counts where the rows represent each plasmid, the columns represent each *k*-mer and the entries represent the *k*-mer counts. Additionally, a merged matrix was generated by combining the *8*-mer count matrix with the genome length, GC content and codon usage information.

In this study, we generated multi-label models that are able to predict multiple hosts per plasmid. Each label corresponds to a plasmid host and encodes a binary value, with “1” corresponding to being a host. These plasmid hosts were predicted at different taxonomy levels such as species, genus, family and order, where we built separate models per taxonomic level. We used random forest to build the classifiers, which provides robust and interpretable predictions based on decision trees and has been explored in many other classification studies (18, 30-32).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Model parameter tuning and validation was performed using the plasmid data, where the data were split into training, testing and hold-out datasets. The training and testing datasets were used for parameter tuning, and the hold-out dataset used for monitoring possible overfitting. Random forest was implemented using *ensemble.RandomForestClassifier* from the Python Scikit-learn package (version: 0.20.4) (33). The model parameters were tuned in the five-fold cross validation loop using the random grid search method from Scikit-learn that iterated 100 times; n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf and bootstrap were the parameters tuned (Table S1). These parameters were responsible for the number of trees in the forest, the number of features required for the split, the maximum depth of the tree, the minimum number of samples for splitting, the minimum number of samples required for the leaf, and bootstrapping of samples, respectively. Tuning was conducted using an *8*-mer matrix at the genus level and then applied to the other taxonomic levels and *k*-mer sizes. The detected optimal parameters were n_estimators = 1,000, max_features = "auto" which is square root of the number of features, max_depth = 50, min_samples_split = 2, min_samples_leaf = 1, and bootstrap = False. The class_weight parameter was set to “balanced” to weight the inputs based on the class frequencies to prevent the biased predictions due to the imbalanced classes. The random forest model was utilized with the *multiclass.OneVsRestClassifier* from the Python Scikit-learn (version: 0.20.4) package (33) which fits one label at a time and improves the interpretability of the models.

Using the tuned parameters, the random forest model was trained and tested five times using the *k*-fold cross-validation method, where different datasets were tested each time. The ensembled cross-validation model was applied on the hold-out dataset which was not part of the training or testing. Model performances were measured using the area under curve (AUC), macro F1 score, Matthews correlation coefficient (MCC), and the confusion matrix using Scikit-learn (33). Sensitivity and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

specificity were also calculated from confusion metrices to measure the ability of model to identify hosts. The class probability threshold of 0.5 was applied to calculate the performances of macro F1, MCC, and confusion matrix. Since it is a multi-label problem, all of the predictions for all possible labels were collected into one data container, and one prediction performance was calculated per model instead of the number of labels. The test and hold-out dataset performances were reported. The test performances were calculated by averaging five model performances from the cross-validation, and reported with standard deviations. The hold-out performances were the ensemble model performances from averaging the cross-validation model predictions.

To test whether the plasmid host model predictions were significantly different, a t-test was performed using *stats.ttest*_ind from SciPy (version: 1.2.2) (34). Moreover, possible correlations were detected using Spearman's correlation coefficient using *stats.spearmanr* from SciPy (version: 1.2.2).

Clustering plasmids

Since similar plasmids are likely to be hosted by the same organisms, we clustered the plasmids based on *k*-mer sequence similarity using KMA index (version: 1.3.9) with the following parameters: -k16, -Sparse and -NI (24). KMA clusters the sequence for a given similarity threshold using *16*-mers and the Hobohm-1 algorithm (35). We clustered the plasmids using three different *k*-mer query and template similarity thresholds at 90%, 80% and 50%. By dividing the clusters into training, testing and hold-out sets; similar plasmids were kept in the same partitions. This forced the models to learn sequence characteristics spanning larger genetic distances and was intended to help improve the generalizability of the models.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Random fragments

Partial sequences might be informative for predicting hosts and better reflect actual data in incomplete plasmid assemblies. Therefore, random fragments of 500, 1,000, and 1,500 nucleotides were sub-sampled from each plasmid sequence to build prediction models from the partial sequences. The sub-sampling process was repeated randomly ten times for each plasmid. Plasmids shorter than the given fragment size were excluded from the study and separate models were built per fragment size. Matrix files and models were constructed as described above, using *k*-mers that were 5 and 8 nucleotides in length. *10*-mers were not utilized due to the heavy computational requirements.

Validation of the plasmid host prediction models

The plasmid host models that were trained with the PATRIC plasmids at four different taxonomic levels were validated using plasmids from the National Center for Biotechnology Information (NCBI) Reference Sequence database (RefSeq) (36). A total set of 30,349 NCBI plasmids were downloaded from the NCBI RefSeq database and filtered. NCBI offers a larger plasmid collection than PATRIC, yet some of the plasmids are identical. Therefore, only the plasmids that had not yet been integrated into PATRIC as of January 2021, *i.e.*, plasmids that are only present in the NCBI RefSeq database were included. Further, we eliminated duplicates from the NCBI validation dataset by comparing *k*-mer counts, and filtered based on the source organism, completeness and NCBI's automatic taxonomy check. Moreover, plasmids with labels that are not included in the PATRIC training data were further removed from the NCBI validation data. The remaining plasmids with

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

species level host information recognized by NCBI Taxonomy were tested against the plasmid host models that were trained on the PATRIC collection. The validation performances were reported in AUC, macro F1, MCC and the confusion matrix. The class probability threshold of 0.5 was applied to calculate the performances of macro F1, MCC and confusion matrix.

Comparison to PlasmidFinder

We compared our species model performance to PlasmidFinder for the *Enterobacteriaceae* species that are present in the PATRIC hold-out dataset (3, 37). Moreover, the hold-out plasmids that present in the PlasmidFinder database were excluded from this comparison. The PlasmidFinder tool (version: 2.1.1) was performed with the default parameters using the PlasmidFinder database downloaded in July 2021.

RESULTS

Plasmid host prediction performances for the PATRIC hold-out dataset

In order to develop models for predicting the host organisms of plasmids, a total of 8,519 plasmids with at least species level host information were downloaded and curated from the PATRIC database and included in this study (Table S2). These plasmids originate from 359 species belonging to 174 genera, 93 families, and 50 orders (Figure S1, Table S3). Most of the plasmids in the collection come from the orders Enterobacterales, Bacillales, and Lactobacillales, which comprise 55.6% of the hosts in the data set (Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To predict the taxonomic label of the host organism, machine learning models were trained using nucleotide *k*-mer counts from the plasmids. The predictions were carried out using *5*, *8* and *10*-mers, since the short and long *k*-mers might provide different types of information to the models. For example, *5*-mers do not usually appear in the plasmid genome uniquely, and instead provide the models with information regarding the profile of oligonucleotide frequencies for each plasmid. On the other hand, the longer *k*-mers, such as *8*- and *10*-mers, usually occur uniquely in a given plasmid, and offer counts of unique sub-sequences. Moreover, *k*-mer distributions are subject to changes based on the sequence size.

Using each *k*-mer size, random forest-based classifiers were built to predict host taxonomy from order to species levels. The model based on the *5*-mer counts has 0.655 MCC for predicting the plasmid host species, and this was moderately higher, 0.662 and 0.680 MCC, for *8*-mers and *10*-mers, respectively (Figure 2, Table S4-S6). At the order level, the model performances achieved 0.899 MCC for *5*-mers, 0.867 MCC for *10*-mers and *8*-mers (Figure 2, Table S4-S6).

By increasing the *k*-mer size from five to ten, the prediction performances increased 3.8% in MCC at the species level but decreased 3.6% at the order level. Although the fluctuations in the performances are not significant according to the paired t-test (p-values [0.404, 0.883] > significance threshold 0.05). To limit computational needs, we used the *8*-mers, but not *10*-mers, to build input matrices for all sub-sequent analyses. Overall, the plasmid host prediction models have low sensitivity (true positive rate) and high specificity (true negative rate) where the lowest sensitivity was detected at the species level compared to other taxonomy levels where sensitivity falls into the range between 0.493 and 0.761.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The number of the false negative predictions increased inversely with the presence of the hosts in the input data (Figure 3). Moreover, this correlation was significant at the species level (Spearman's correlation coefficient of 0.545, p-value 0.00 < significance threshold 0.05). These findings suggest that the host classification becomes more challenging at the species level, and the model performances improve proportionally to the host representations in the training data. In addition to the false negatives, false positive predictions made by the model (Figure S3-S4) were frequently phylogenetically close to the actual hosts. For instance, the model frequently predicted *Escherichia coli* hosts instead of the *Salmonella enterica* and *Klebsiella pneumoniae,* where all of them belong to the *Enterobacteriaceae.*

In an attempt to improve the *8*-mer model performances, we combined the *k*-mer frequencies with the information in nucleotide compositions of plasmid sequences including, plasmid size, GC content and codon usage. These additional features yielded an approximately 0.6%-1.6% increase in the MCCs of the models (Figure S2, Table S7); however, this improvement was not significant according to the paired t-test (p-value 0.892 > significance threshold 0.05). Therefore, the following analyses were carried out without these additional features.

To understand the impact of plasmid sequence similarity on the model performances, the plasmid genomes were clustered based on the *k*-mer similarity using KMA. The plasmids belonging to the same cluster at a given *k*-mer similarity threshold were kept in the same training, testing or hold-out dataset. When the *k*-mer similarity decreased to 80%, thus making the clusters more diverse, the model performances decreased in MCC between 7.7% to 29.8% depending on the taxonomic level (Figure 4, Table S8). The performance decrease shows that the similarity between the training and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

testing data has an effect on the host predictions, especially at the lower taxonomic levels, although, the model can still be generalized to distant sequences.

Plasmid host predictions with random fragments

Due to the fragmented nature of plasmid assemblies that results from the difficulty in assembling plasmids from the short reads, we wanted to develop random forest models that can make predictions from incomplete sequences. To do that, we trained and tested our plasmid host prediction models with random fragments of plasmid sequences. Fragments of 500, 1,000 or 1,500 nucleotides were randomly sampled from each assembled plasmid sequence over ten rounds. By sampling multiple times, we attempted to introduce various regions of the plasmid sequences to the models. The fragment model that was trained with the 500 nucleotide fragments using *5*-mers reached 0.426 MCC for the species model and 0.674 for the order model (Figure 5, Table S9). When the same fragments were sub-sampled into *8*-mers, the species level model had MCCs of 0.489 and 0.686 MCC for the species and order levels, respectively (Figure 5, Table S10). By increasing the fragment size from 500 to 1,000 nucleotides, the model performances increased 8.2%-10.7% in MCC with the *5*-mers and 6.3%-10.1% in MCC with the *8*-mers (Figure 5, Table S9-S10). When the fragment size increased from 1,000 nucleotides to 1,500 nucleotides, the model performances increased 4.8%-5.1% in MCC with the *5*-mers and 3.3%-1.7% in MCC with the *8*-mers (Figure 5, Table S9-S10). The fragment models reached their highest performances using 1,500 nucleotide fragments and *8*-mers as the features, where the MCCs were 0.537 and 0.768 for the order and species model, respectively (Figure 5, Table S10).

Validation of the plasmid host prediction model with the NCBI validation dataset

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To validate the plasmid host prediction models, we used plasmids in the NCBI RefSeq collection that are not present in our training, test or hold-out datasets. Overall, 7,670 bacterial plasmid sequences with taxonomic metadata were included in this analysis (Table S11-S12). As in the PATRIC database, the NCBI validation data is also dominated by the few major orders such as Enterobacterales, Lactobacillales and Pseudomonadales, which make up approximately 76% of the data set (Figure 6).

When the whole model (trained with *8*-mers of the PATRIC training set) was tested with the NCBI validation data, the ratio of the correct and wrong predictions was shown in Figure 7. Our plasmid host prediction model has relatively low sensitivity (0.483) and a high specificity (1.0) at the species level (Table S13), similar to the results shown above. Moreover, when the NCBI validation data were tested with the random model generated by shuffled labels, the model performance dropped to 0.028 MCC at the species level. This suggests even though the sensitivity is low, the model has adequate generalizability which is far from being random.

Because the NCBI collection contained many short plasmid sequences, we filtered it based on the sequence size. Overall, plasmid sequences equal or greater than 5,000 bp performed 43% better than plasmid sequences less than 5,000 bp in terms of MCC at the species level. However, this performance gap reduced to 1% at the order level (data not shown). This means that the plasmid host range model accuracy improves with longer plasmid sequence length at lower taxonomic levels.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Similarly to the predictions for the hold-out dataset, the plasmid host prediction model predicted additional hosts for 499 plasmids in the NCBI validation dataset (Figure S5). Furthermore, the model frequently erroneously predicted *Escherichia coli* as being the host instead of its close relatives, *Salmonella enterica* and *Klebsiella pneumonia*, both in the hold-out and the validation results.

The fragment-based models were also validated using the NCBI dataset. The 7,670 NCBI plasmids were randomly sub-sampled into 500, 1,000, and 1,500 nucleotide fragments, and each plasmid was randomly sampled ten times per fragment size. Similar to the PATRIC results, the fragment models reached the best performances (0.485 MCC at the species level and 0.778 MCC at the order level) for the NCBI validation data with the 1,500 nucleotides fragment size and *8*-mers (Figure 8, Table S14-S15).

Comparison to PlasmidFinder

The PlasmidFinder tool uses an alignment-based strategy to identify plasmid sequences, and can often provide host information when it is available. We used 391 *Enterobacteriaceae* plasmids in the PATRIC validation data that were not already part of the PlasmidFinder database to compare the output of PlasmidFinder and our machine learning models. Overall, PlasmidFinder correctly identified 304 of the sequences and did not predict anything for 87 sequences. Our whole-plasmid-based *8*-mer model successfully classified 229 of the plasmid hosts, nothing predicted 121 and falsely predicted 41 of them at the species level. We then randomly sampled the 391 plasmids into 1,500 nucleotide fragments and compared PlasmidFinder with our *8*-mer based model based on 1,500 nucleotide fragments. Overall, PlasmidFinder is able to identify 228 out of 3,910 fragmented

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

plasmid sequences. However, none of the returned matches contained plasmid host information. Our fragment-based model predicted a host for 1,927 of the fragmented plasmid sequences correctly, nothing predicted 1,309 and falsely predicted 674 of them. Compared to our machine learning model, the alignment-based PlasmidFinder tool provides accurate predictions for the *Enterobacteriaceae* species when a sequence match is available and plasmids sequences are mostly complete. When the plasmids are fragmented, the machine learning strategy becomes more advantageous.

The web-server

The plasmid host prediction models that were trained using *8*-mers from whole plasmid sequences can be used online on the Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/PlasmidHostFinder/). This web-server tool accepts one FASTA file at a time and provides an output file containing the predicted plasmid host range at the selected taxonomic level from species to order. The web-server tool enables two model options, fast and slow, with various class thresholds. The slow model uses all five cross-validation models to make a final decision on the plasmid host range. The fast mode uses only the first cross-validation model out of five to predict the plasmid host range. Therefore, one can expect to obtain more confident predictions with the slow model.

DISCUSSION

In this study, we built random forest models that can predict plasmid hosts and host-ranges at taxonomic levels between species and order; these models achieved accuracies from 0.662 to 0.867

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

MCC. The model performs better at higher taxonomic levels, with 'order' level being the best. We observed that the *k*-mer size does not have a significant influence on the prediction performances. Among the three *k*-mer sizes, we chose to build our prediction models with *8*-mers since it provides robust predictions at all taxonomic levels with less computational effort than *10*-mers. Moreover, we tried to improve the host range predictions with the additional genome features such as plasmid size, GC content and codon usage but the increase in the prediction performances was negligible. We validated our models using an independent dataset from the NCBI RefSeq. These performances were comparable with our previous test and validation results. In addition, to assess the utility of this approach with partially assembled plasmid sequences, we generated models for 500, 1,000, and 1,500 nucleotide fragments, and even the smallest fragments of 500 nucleotides have sufficient information for the identification of plasmid hosts.

Machine learning

We observe that the robustness of the models is dependent on the quantity, quality and accuracy of the input and output data. In this study, the plasmid host prediction models might suffer from incomplete metadata despite our best efforts. The plasmid data and corresponding plasmid hosts were retrieved from the PATRIC database. However, the PATRIC dataset is likely to contain some plasmids with incomplete host range information. This issue might have an effect on robustness of the models, but is most likely to have a minor impact due to the relatively large input dataset. Nevertheless, some of the false positive predictions might be the consequence of incomplete metadata. Some of the other false positives might potentially be new discoveries relating to plasmid transmission in diverse hosts, although this theory should be validated experimentally.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Plasmid genomes are extremely plastic (38). Accessory genes vary in their presence or absent from the plasmids, which makes plasmid host prediction a complicated task. In order to understand the impact of the genome similarity on the plasmid host model learning, we clustered the plasmids for a given similarity threshold. By keeping the similar plasmids in the same training, test or hold-out datasets, the learning from the sequence similarity was minimized since the similar plasmids tend to have the same hosts. This clustering approach caused less accurate results than the baseline model. These results suggest that sequence similarity has an impact on the model learning. Therefore, to boost the model performances, the training data should be updated regularly to increase the input diversity when more plasmid data is available. In addition to the sequence similarity, host related signals from the relatively conserved regions of the plasmid sequences such as *rep* or *mob* genes are likely learned from the model. Further analysis of the top model features may help to validate or elucidate genetic features involved in transmission, especially in less studied taxonomic groups.

The model performances were evaluated using several performance measurements including AUC, MCC, and macro F1. The AUC and MCC performances were not always correlated and caused different conclusions in some cases such as in the random forest model with the clustered plasmids. The reason for this discrepancy may be the applied class thresholds. AUC uses a range of thresholds to measure the model performances and does not require a defined class threshold. In contrast to AUC, the MCC and macro F1 calculation require predictions instead of probabilities. Therefore, a defined class probability threshold is needed for converting probabilities to predictions. This threshold was set to 0.5 for all the models. But, this threshold might not be the ideal threshold for some of the models, particularly for the imbalanced classes (39). For instance, the species level prediction model has a lower sensitivity (0.493) compared to its specificity (1.0). In other words, the model failed to predict some of the hosts (Figure 3,7), and the majority of the failed predictions

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

were the result of having no positive class predicted for the tested plasmid due to no predictions being above or equal to the class probability threshold of 0.5. Therefore, adjusting the class probability threshold could be a way to improve the model.

False positives or unknown hosts

The machine learning models have the potential for discovery of unknown correlations between the input features and predicted phenotypes. For example, in previous studies, novel AMR genes were reported using the machine learning models (18, 19). In our case, machine learning might be useful for discovering unknown plasmid hosts. We explored the false positives as in: 1) the model was not able to predict the actual hosts, but predicted false positives (Figure S3), 2) the model predicted multiple hosts including the actual hosts and false positives (Figure S4). These cases should be investigated further as these could happen due to two reasons: the model might pick up noise, or the falsely predicted host might actually be a host in nature. Thus, a portion of the false positives might be the actual hosts which are not discovered before, but machine learning gives the opportunity for it *in silico*. To prove that they are potential hosts would require *in vitro* experiments to test the stability of the plasmid in these bacteria.

Fragments

The fragment-based model performances vary based on the fragment and *k*-mer sizes. We obtained the best performances for the hold-out dataset with the 1,500 nucleotide fragments using *8*-mers. The fragment size and model performances changed proportionally because the longer fragments are providing more information to the models. This correlation between the model performances

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and fragment size might be the consequence of the mosaic nature of plasmids. Genes located on plasmids could originate from different organisms and random sampling of these acquired genes might cause false predictions. Moreover, as the plasmids were not aligned prior to the fragmentation, the genetic content of fragments that sub-sampled from different plasmids did not match. Therefore, we expect the model learning the fragment structures instead of the unique patterns.

Conclusion

We built random forest models and incorporated them in PlasmidHostFinder tool to detect plasmid hosts and host-ranges at various taxonomic levels from species to order with the performance of 0.662 MCC to 0.867 MCC. PlasmidHostFinder can detect a diverse range of hosts for 359 species, 174 genera, 93 families and 50 orders with high accuracy in spite of the mosaic, diverse nature and genetic plasticity of plasmids. The approach described in this study helps to fill a gap in our ability to predict plasmid hosts, particularly in understudied taxa, or when plasmid sequences are fragmented.

REFERENCES

1. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nature Reviews Microbiology doi:10.1038/s41579-020-00497-1.

2. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M, Peto T, Crook D, Walker AS, Woodford N, Anjum MF, Stoesser N. 2017. Ordering the mob: Insights into replicon and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid 91:42-52.

3. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895-903.

4. Lozano C, García-Migura L, Aspiroz C, Zarazaga M, Torres C, Aarestrup FM. 2012. Expansion of a plasmid classification system for Gram-positive bacteria and determination of the diversity of plasmids in Staphylococcus aureus strains of human, animal, and food origins. Appl Environ Microbiol 78:5948-55.

5. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, Crook D, Woodford N, Walker AS, Phan H, Sheppard AE. 2017. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Front Microbiol 8:182.

6. San Millan A, MacLean RC. 2017. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. Microbiol Spectr 5.

7. Unterholzner SJ, Poppenberger B, Rozhon W. 2013. Toxin-antitoxin systems: Biology, identification, and application. Mob Genet Elements 3:e26219.

8. Bello-López JM, Cabrero-Martínez OA, Ibáñez-Cervantes G, Hernández-Cortez C, Pelcastre-Rodríguez LI, Gonzalez-Avila LU, Castro-Escarpulli G. 2019. Horizontal Gene Transfer and Its Association with Antibiotic Resistance in the Genus *Aeromonas* spp. Microorganisms 7:363.

9. Jain A, Srivastava P. 2013. Broad host range plasmids. FEMS Microbiology Letters 348:87-96.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

10. Suzuki H, Yano H, Brown CJ, Top EM. 2010. Predicting Plasmid Promiscuity Based on Genomic Signature. Journal of Bacteriology 192:6045-6055.

11. Robertson J, Bessonov K, Schonfeld J, Nash JHE. 2020. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. Microb Genom 6.

12. Suzuki H, Brown CJ, Top EM. 2018. Genomic Signature Analysis to Predict Plasmid Host Range, p 458-464. *In* Wells RD, Bond JS, Klinman J, Masters BSS (ed), Molecular Life Sciences: An Encyclopedic Reference doi:10.1007/978-1-4614-1531-2_574. Springer New York, New York, NY.

13. Suzuki H, Sota M, Brown CJ, Top EM. 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res 36:e147.

14. Norberg P, Bergstrom M, Jethava V, Dubhashi D, Hermansson M. 2011. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. Nat Commun 2:268.

15. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46:e35.

16. Redondo-Salvo S, Bartomeus-Penalver R, Vielva L, Tagg KA, Webb HE, Fernandez-Lopez R, de la Cruz F. 2021. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics 22:390.

17. Xu C, Jackson SA. 2019. Machine learning and complex biological data. Genome Biology 20:76.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

18. Aytan-Aktug D, Clausen PTLC, Bortolaia V, Aarestrup FM, Lund O. 2020. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. mSystems 5:e00774-19.

19. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, Heckmann D, Anand A, Yang L, Nizet V, Monk JM, Palsson BO. 2018. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Nature communications 9:4306-4306.

20. Ruppe E, Ghozlane A, Tap J, Pons N, Alvarez AS, Maziers N, Cuesta T, Hernando-Amado S, Clares I, Martinez JL, Coque TM, Baquero F, Lanza VF, Maiz L, Goulenok T, de Lastours V, Amor N, Fantin B, Wieder I, Andremont A, van Schaik W, Rogers M, Zhang X, Willems RJL, de Brevern AG, Batto JM, Blottiere HM, Leonard P, Lejard V, Letur A, Levenez F, Weiszer K, Haimet F, Dore J, Kennedy SP, Ehrlich SD. 2019. Prediction of the intestinal resistome by a three-dimensional structure-based method. Nat Microbiol 4:112-123.

21. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, Xia F, Xie D, Yoo H, Stevens R. 2020. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic acids research 48:D606-D612.

22. Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24.

23. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020.

24. Clausen PTLC, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 19:307.

25. Hallgren MB, Overballe-Petersen S, Lund O, Hasman H, Clausen P. 2021. MINTyper: an outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads. Biol Methods Protoc 6:bpab008.

26. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 47:W256-W259.

27. Sievers A, Bosiek K, Bisch M, Dreessen C, Riedel J, Froß P, Hausmann M, Hildenbrand G. 2017. K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. Genes 8:122.

28. Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33:2759-2761.

29. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

30. Pataki BÁ, Matamoros S, van der Putten BCL, Remondini D, Giampieri E, Aytan-Aktug D, Hendriksen RS, Lund O, Csabai I, Schultsz C, Matamoros S, Janes V, Hendriksen RS, Lund O, Clausen P, Aarestrup FM, Koopmans M, Pataki B, Visontai D, Stéger J, Szalai-Gindl JM, Csabai I, Pakseresht N, Rossello M, Silvester N, Amid C, Cochrane G, Schultsz C, Pradel F, Westeel E, Fuchs S, Kumar SM, Xavier BB, Ngoc MN, Remondini D, Giampieri E, Pasquali F, Petrovska L, Ajayi D, Nielsen EM, Trung NV, Hoa NT, Ishii Y, Aoki K, McDermott P, group SCM-A. 2020. Understanding and predicting ciprofloxacin minimum

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

inhibitory concentration in *Escherichia coli* with machine learning. Scientific Reports 10:15026.

31. Breiman L. 2001. Random Forests. Machine Learning 45:5-32.

32. Sarica A, Cerasa A, Quattrone A. 2017. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience 9.

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Louppe G, Prettenhofer P, Weiss R. 2012. Scikit-learn: machine learning in python. arXiv. arXiv preprint arXiv:12010490.

34. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17:261-272.

35. Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. Protein science : a publication of the Protein Society 1:409-417.

36. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-45.

37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-10.

38. Kostlbacher S, Collingro A, Halter T, Domman D, Horn M. 2021. Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria. Curr Biol 31:346-357 e3.

39. Saito T, Rehmsmeier M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10:e0118432.

DATA AVAILABILITY

The Python 2.7.15 scripts that used in this study are available on Bitbucket (https://bitbucket.org/deaytan/plasmid-host-prediction/src/master/). The web-server is available on Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/PlasmidHostFinder-1.0/). All the PATRIC and the NCBI RefSeq sequences and corresponding metadata can be accessed through the PATRIC (https://www.patricbrc.org) and NCBI (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/) resources, respectively.

ACKNOWLEDGMENTS

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We would like to thank Frederik Teudt for his patiently and carefully testing the PlasmidHostFinder tool.

This work was funded by the Novo Nordisk Foundation (grant NNF16OC0021856: Global Surveillance of Antimicrobial Resistance awarded to FMA and OL).

The funding bodies did not play any role in the design of the study or writing of the manuscript, nor did they have any influence on the data collection, analysis or interpretation of the data, or the results.

The authors declare no conflicts of interest.

FIGURES

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

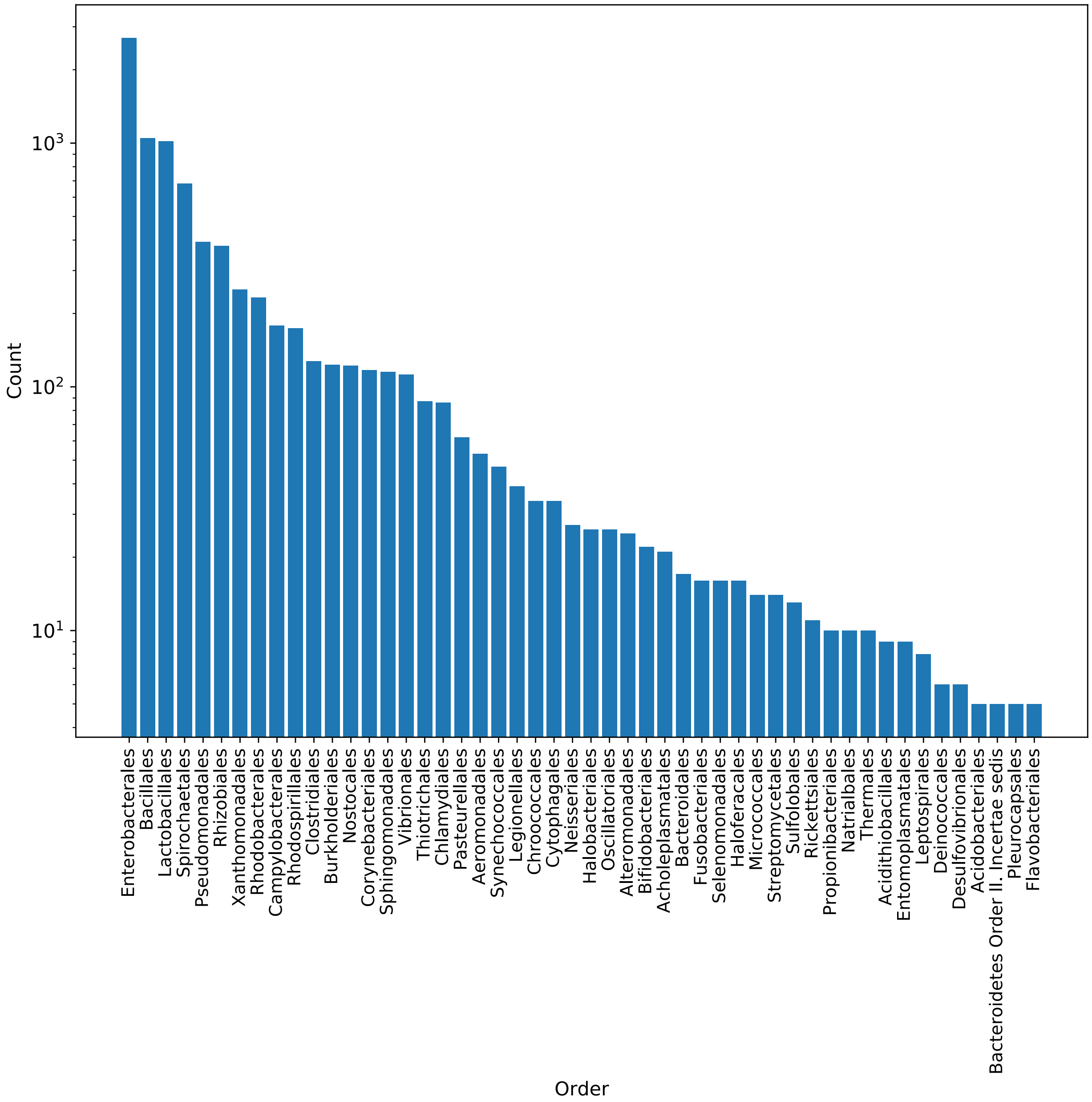


Figure-1: The plasmid host distribution at the order level in the PATRIC dataset. The PATRIC plasmid collection was dominated by the Enterobacterales, Bacillales and Lactobacillales orders which make up 55.6% of the plasmid hosts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

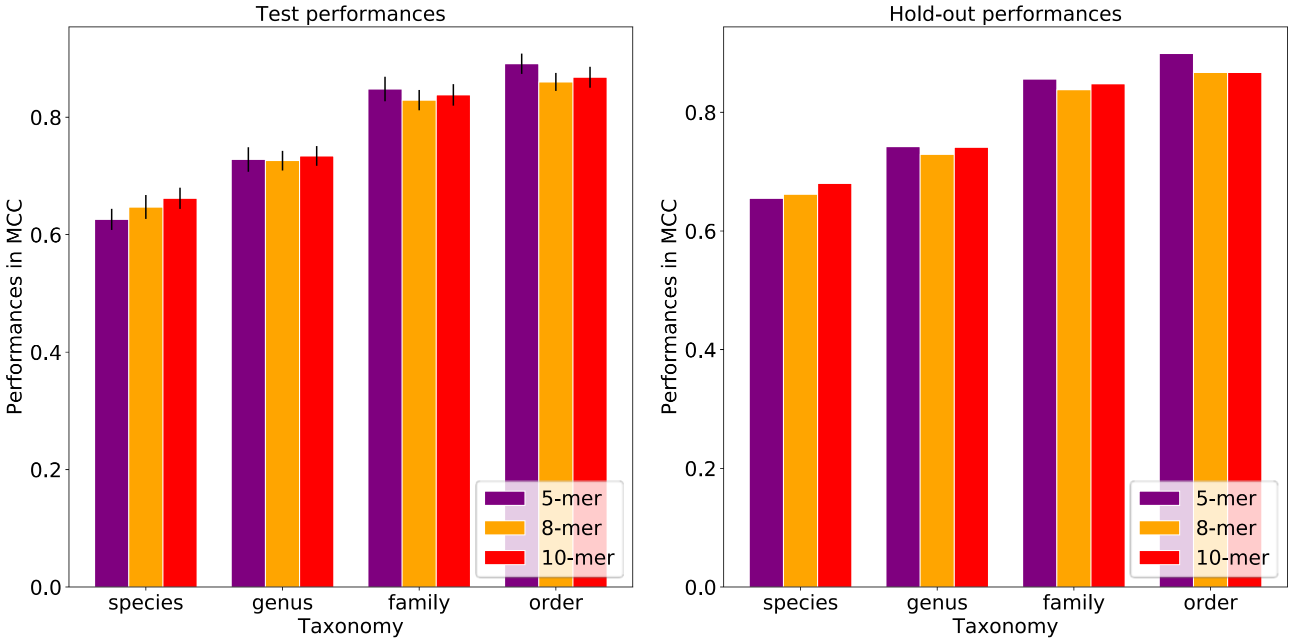


Figure-2: Host prediction performances by *k*-mer size for the test and hold-out data sets. Each bar represents the model performance per taxonomic level. While the test performances were reported with standard deviations, the hold-out performances do not have standard deviations as the five models were combined and a single performance was calculated. The plots show that the prediction performances vary when using different *k*-mer sizes. *5*-mers yield the highest MCC at higher taxonomies, while *10*-mers yield the highest MCC at lower taxonomies. The model performances generally increased from the species to order level for all the *k*-mer sizes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

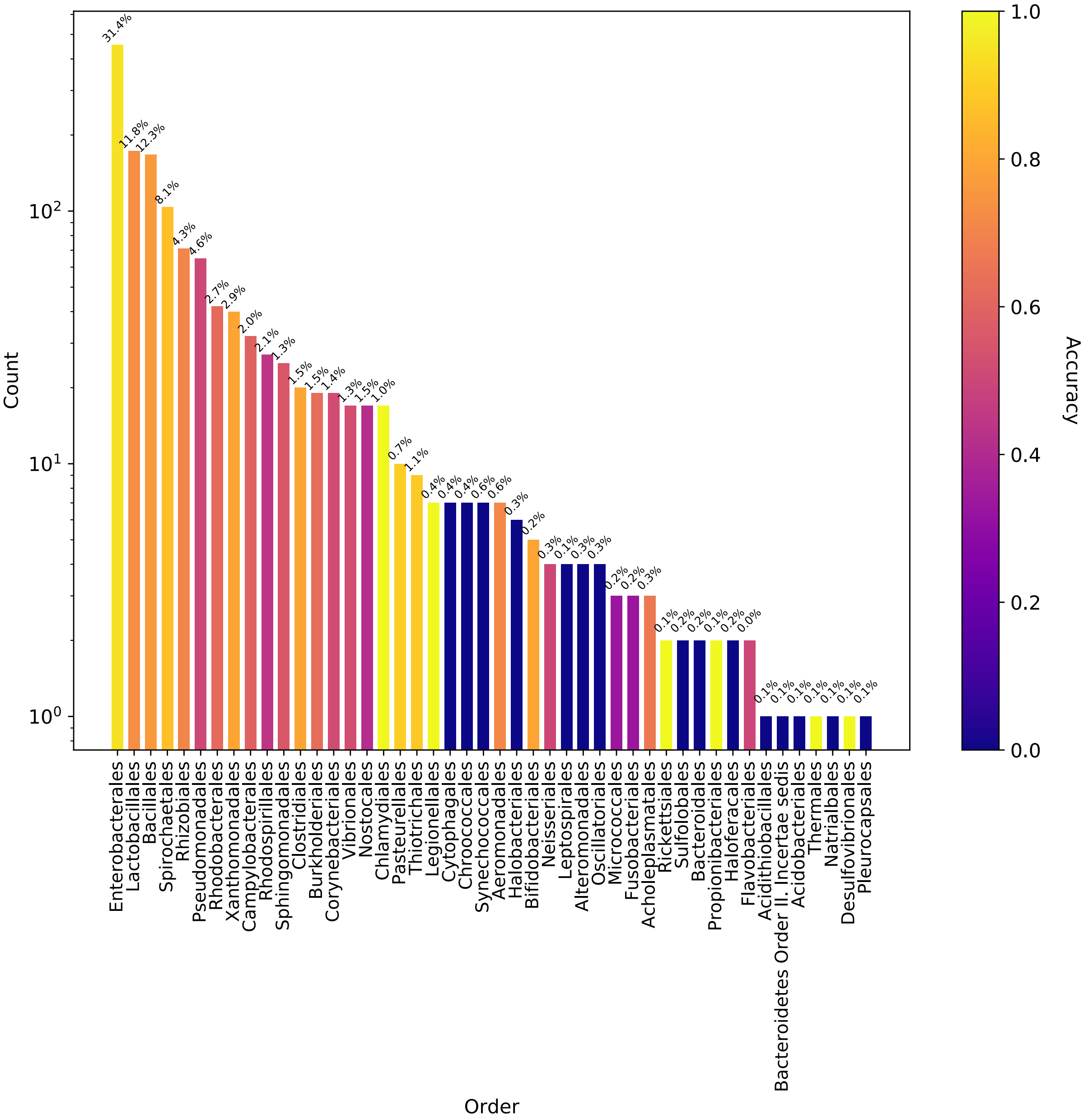


Figure-3: Model accuracy for the PATRIC plasmids tested with the whole model. Each bar shows the number of bacterial orders in the hold-out data and corresponding model accuracy was color coded. The percentage on the top of each bars shows the percentage of bacterial orders in the training data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

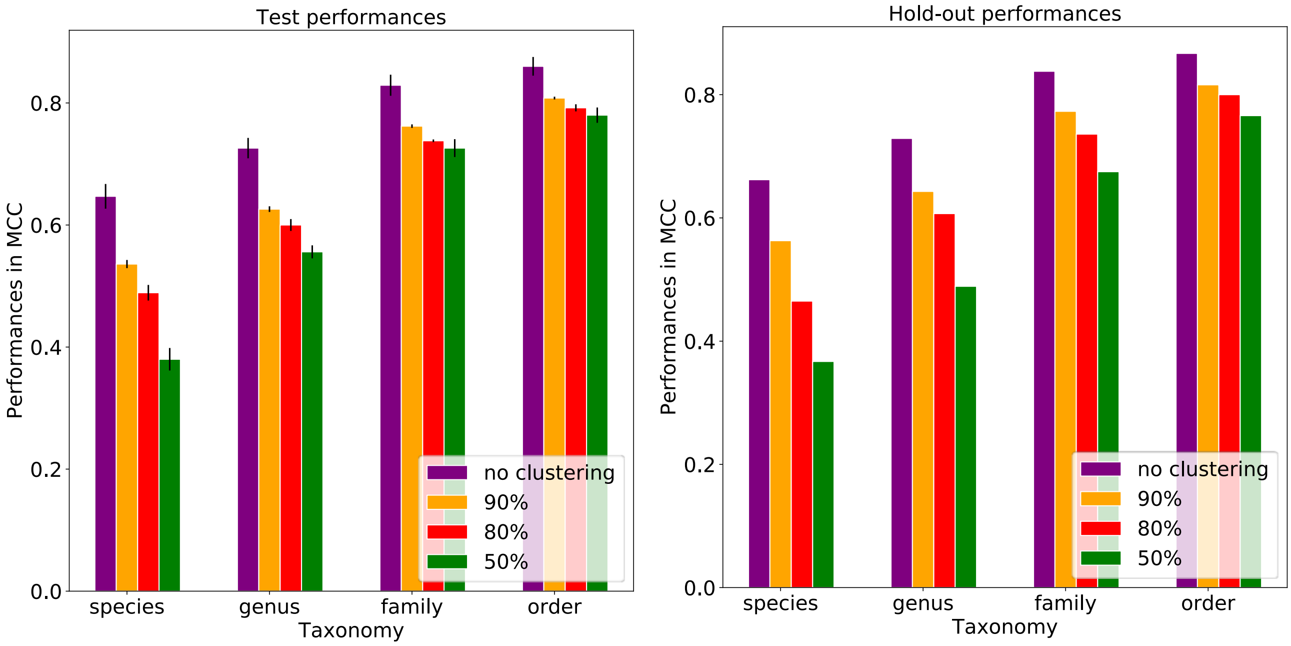
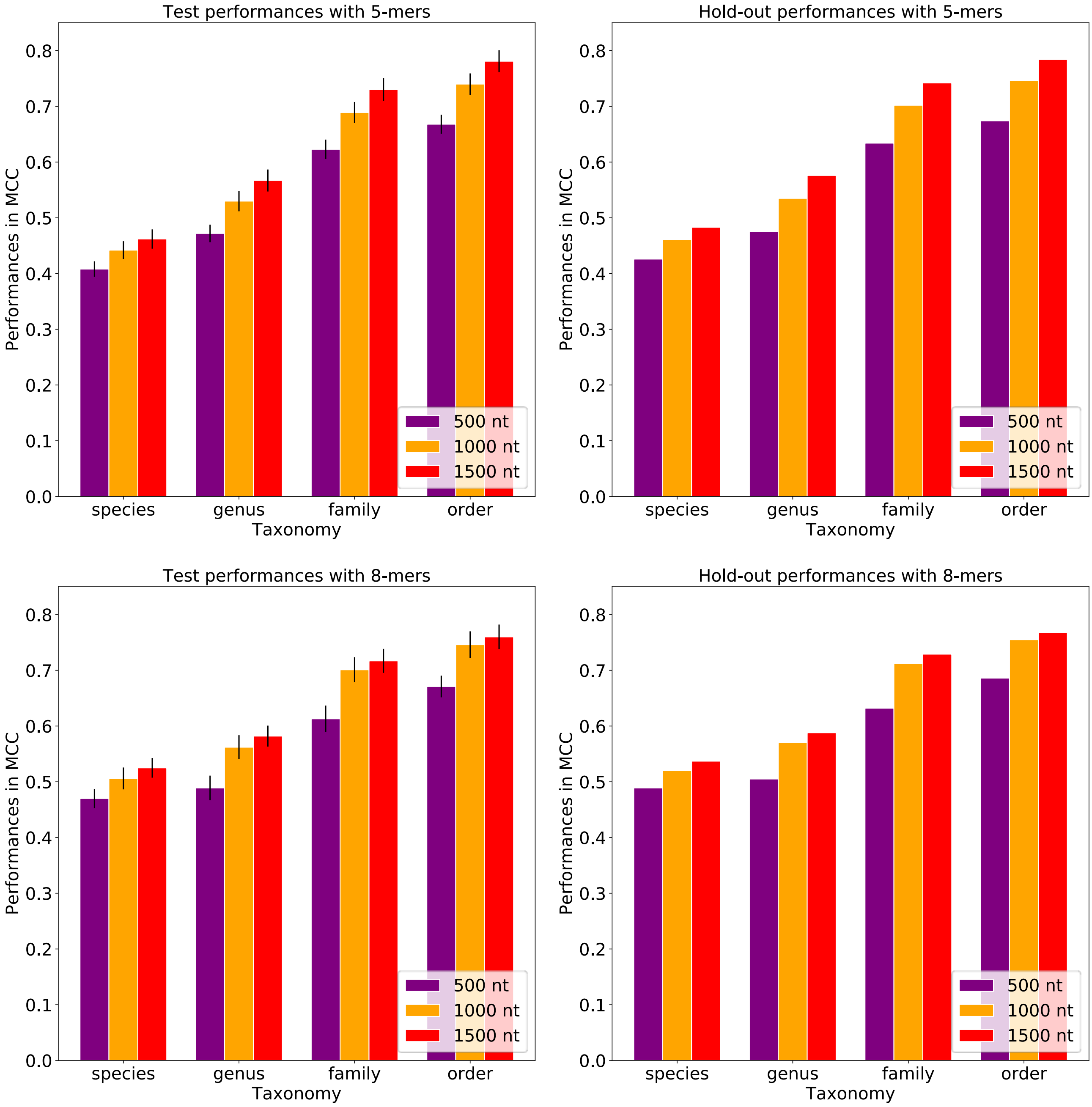


Figure-4: The effects of clustering plasmids at different *k*-mer similarity thresholds on the plasmid host predictions using *8*-mers and different taxonomic levels. Each bar represents the model performance per taxonomic level and each error bar represents standard deviations across folds. The plot shows the influence of plasmid sequence similarity on prediction performances in MCC from the species to order level. The plots suggested that the prediction models pick up sequence similarity mostly at lower taxonomic levels. When the dissimilarity was increased between the training, test and hold-out datasets by applying the 80% *k*-mer similarity threshold; 7.7%-29.8% in MCC performance loss were observed for the hold-out data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



675

676 **Figure-5:** The fragment model performances for the *5*-mer and *8*-mer models. The fragment models were trained with either 500, 1,000, or 1,500 nucleotide (nt) fragments that were sub-sampled from the PATRIC plasmids. The bar plots show the test and hold-out performances for the *5*-mers and *8*-mers in MCC. The error bars represent standard deviations. The best performing model was trained with the 1,500 nucleotide fragments using *8*-mers.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

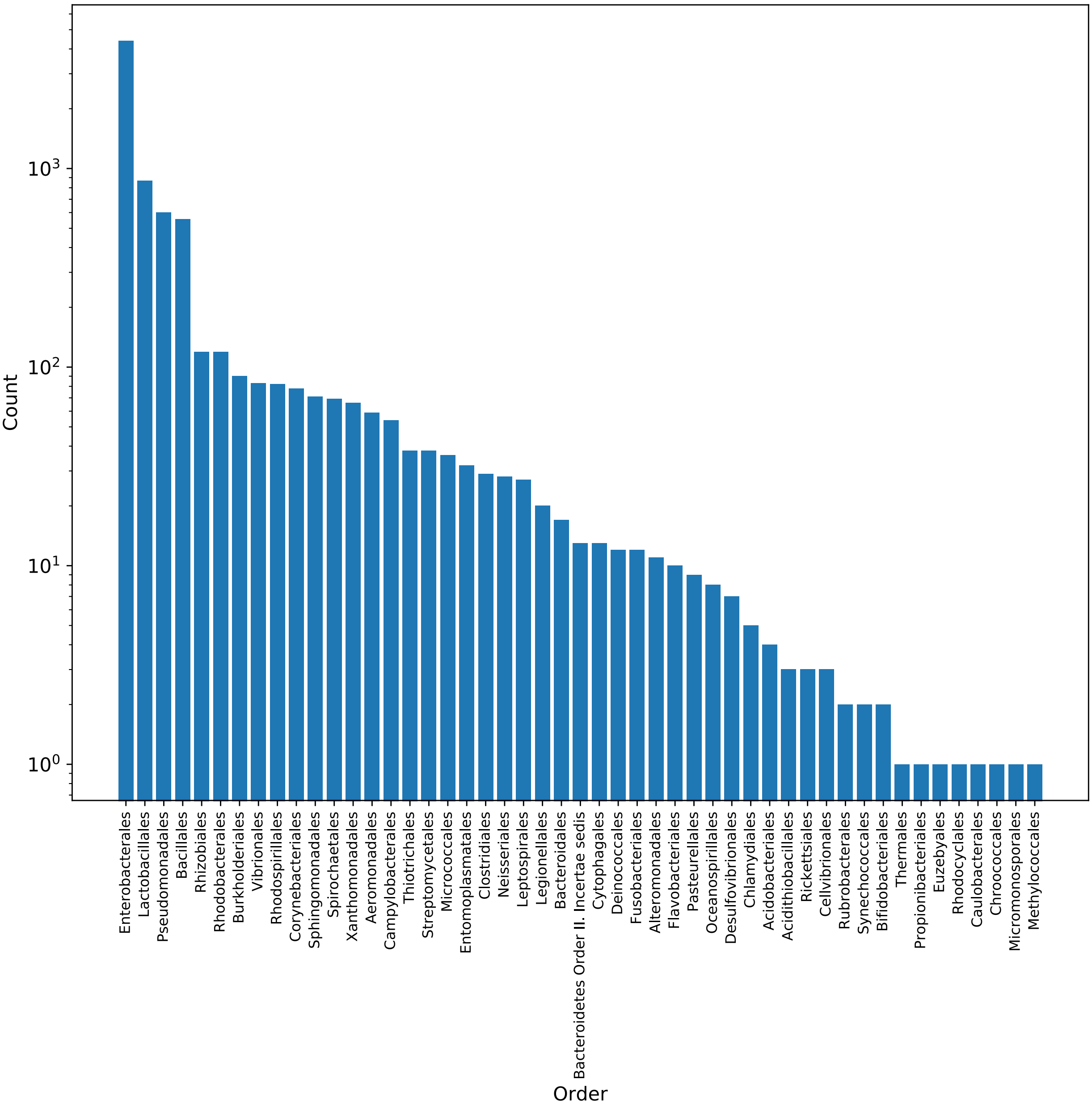


Figure-6: The plasmid host distribution in NCBI validation dataset. The validation dataset was dominated by the Enterobacterales, Lactobacillales and Pseudomonadales, which make up 76% of the NCBI plasmid hosts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

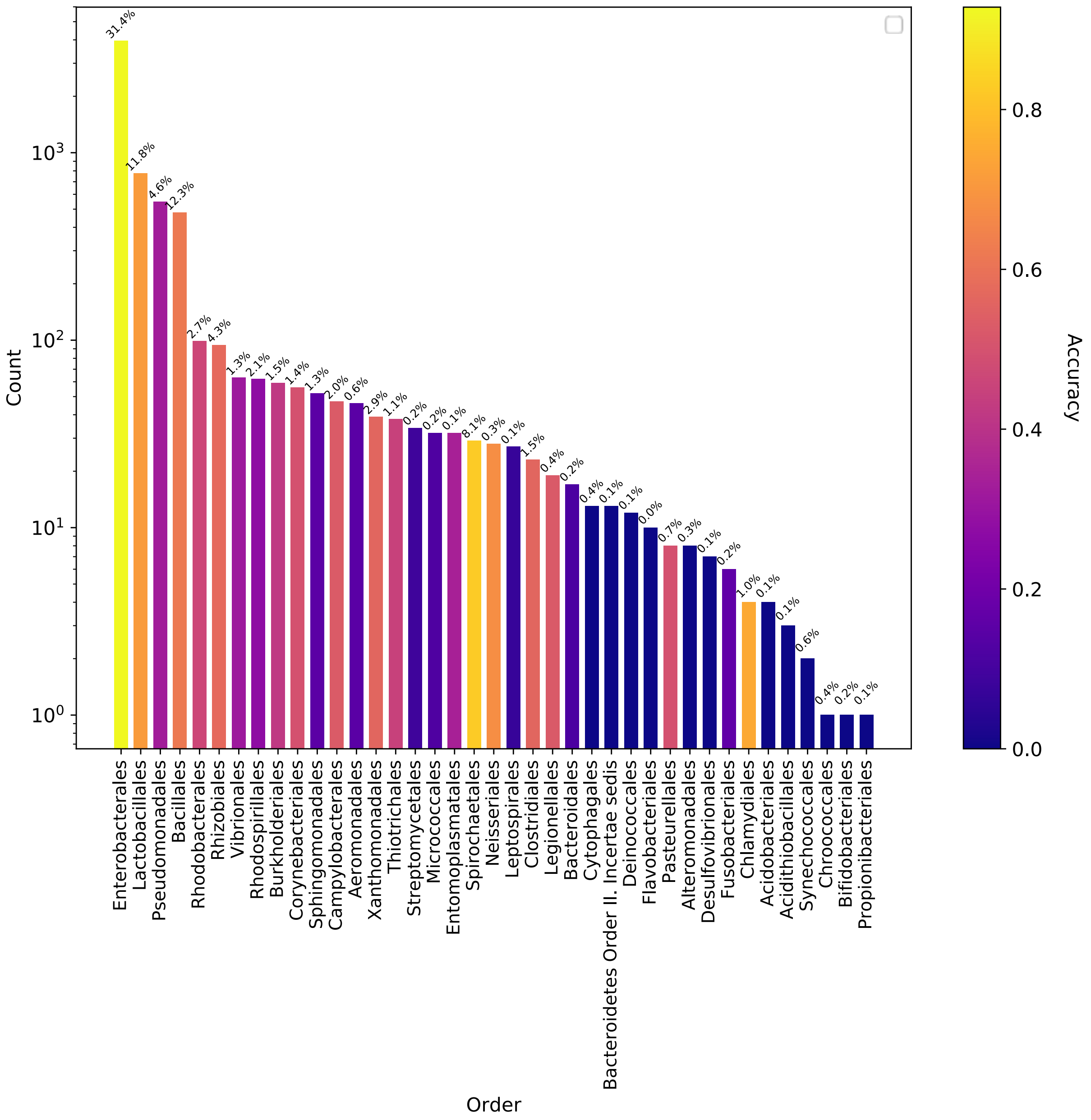
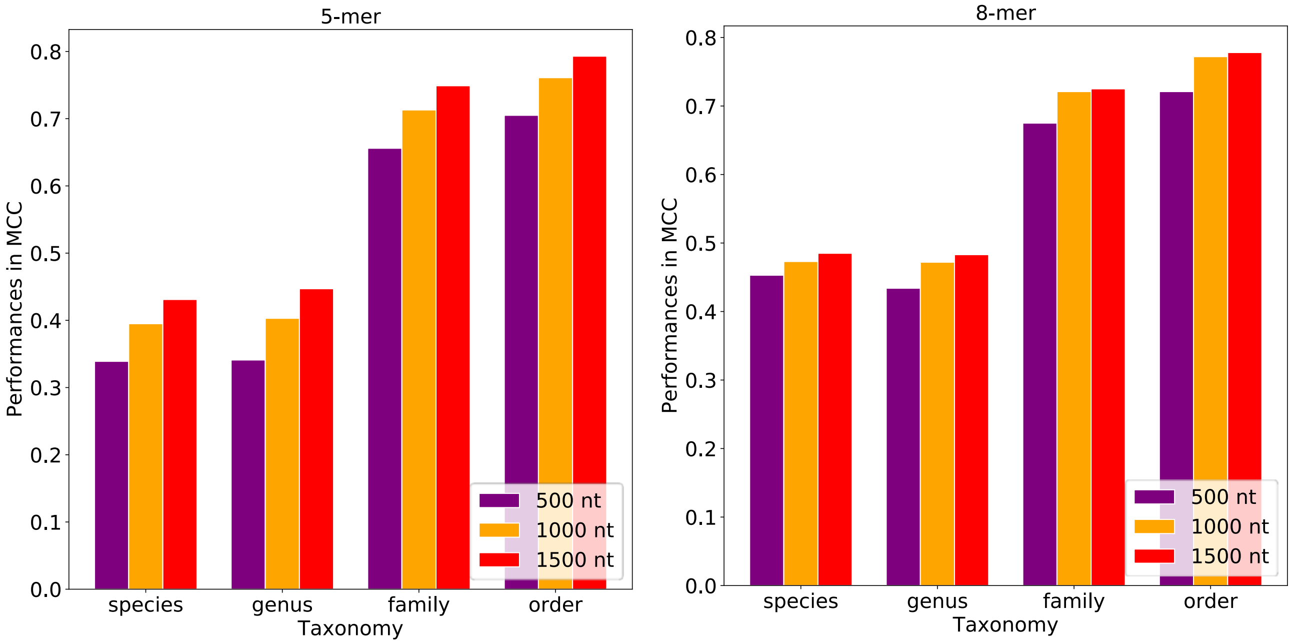


Figure-7: Model accuracy for the NCBI plasmids tested with the whole model that was trained with the PATRIC dataset. Each bar shows the number of bacterial orders in the validation data and corresponding model accuracy was color coded. The plot showed that the accuracy of the models changed roughly according to the availability of the host organisms in the training data which was indicated on top of the bars.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.27.462084; this version posted September 29, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

693 

694 **Figure-8:** The fragment models were validated with the NCBI plasmids. The fragments models that 695 trained with the 500, 1,000 and 1,500 nucleotide (nt) fragments from the PATRIC plasmids were 696 validated with the fragments that sub-sampled from the NCBI plasmids. Similar to the hold-out 697 results, the best performance was obtained with the 1,500 nucleotide fragments and *8*-mers.