


PlasmidHunter: accurate and fast prediction of plasmid sequences using gene content profile and machine learning

Renmao Tian ¹, Jizhong Zhou², Behzad Imanian ^{1,3,*}

¹Institute for Food Safety and Health, Illinois Institute of Technology, 6502 S Archer Rd, Bedford Park, IL 60501, United States

²Institute for Environmental Genomics, Department of Microbiology and Plant Biology, University of Oklahoma, 101 David L Boren Blvd, Norman, OK 73019, United States

³Food Science and Nutrition Department, Illinois Institute of Technology, 10 West 35th Street, Chicago, IL 60616, United States

*Corresponding author. Institute for Food Safety and Health, Illinois Institute of Technology, 6502 S Archer Rd, Bedford Park, IL 60501, United States.

E-mail: bimanian@iit.edu

Abstract

Plasmids are extrachromosomal DNA found in microorganisms. They often carry beneficial genes that help bacteria adapt to harsh conditions. Plasmids are also important tools in genetic engineering, gene therapy, and drug production. However, it can be difficult to identify plasmid sequences from chromosomal sequences in genomic and metagenomic data. Here, we have developed a new tool called PlasmidHunter, which uses machine learning to predict plasmid sequences based on gene content profile. PlasmidHunter can achieve high accuracies (up to 97.6%) and high speeds in benchmark tests including both simulated contigs and real metagenomic plasmidome data, outperforming other existing tools.

Keywords: artificial intelligence (AI); machine learning (ML); plasmid prediction; genomic sequencing

Background

Plasmids are extrachromosomal and transmissible segments of naked, double-stranded DNA that, unlike viruses, replicate autonomously within a host cell. They are common in bacteria, but they are also found in archaea and eukaryota. Plasmids are typically circular and often much smaller than chromosomes, but their sizes vary considerably (from ~1 kbp to >1 Mbp) [1–3].

As agents of horizontal gene transfer (HGT) between bacterial species [4], plasmids spread the traits that might influence the characteristics, survival, and fitness of the hosts, and thus they play an important role in the bacterial evolution and ecology. Plasmids carry non-essential and sometimes beneficial genes

that help their hosts tolerate and survive hostile conditions in a changing environment. For example, a plasmid-borne gene encodes the mercuric reductase that converts toxic Hg²⁺ to a volatile and less toxic metallic Hg⁰ [5]. We now know that plasmids play a key role in spreading the antimicrobial resistance genes (ARGs) among the related bacterial species [6–8], such as those conferring resistance to many commonly used antibiotics such as tetracycline and penicillin or to β -lactams (*bla*) [9] and aminoglycosides (*aad* and *aac*) [10]. In addition to transferring antimicrobial resistance (AMR) to other bacteria, plasmids can transmit and spread other traits [4] such as virulence, toxicity, and pathogenicity to a wider group of bacteria, and consequently, they

Renmao Tian (Tim) is a leading research scientist at the Institute for Food Safety and Health, Illinois Institute of Technology. With interdisciplinary expertise in bioinformatics, machine learning, and microbial genomics, he has published over 70 high-impact papers, garnering 3,100+ citations. Dr. Tian developed widely-used tools like ASAP 2, VBCG, and PlasmidHunter. Dr. Tian's research focuses on computational tools, bacterial pathogenesis, and AI for bioinformatics. His commitment to diversity and innovation drives his groundbreaking contributions to the field.

Jizhong Zhou is a George Lynn Cross Research Professor and Presidential Professor in the Department of Microbiology and Plant Biology, School of Civil Engineering and Environmental Sciences, and School of Computer Science at University of Oklahoma, where he is also Director of the Institute for Environmental Genomics. Dr. Zhou's work is in genomics-enabled microbial environmental sciences. He has advanced experimental and computational metagenomic technologies to address environmental, engineering, and ecological questions. He has led in the elucidation and modeling of microbial feedback mechanisms in response to climate change, anthropogenic pollutions, and environmental gradients. Dr. Zhou earned bachelor's and master's degrees from Hunan Agricultural University in China, and a Ph.D. in Molecular Biology from Washington State University. He is an Adjunct Senior Scientist at Lawrence Berkeley National Laboratory, and a Fellow of the Ecological Society of America, the American Academy of Microbiology, International Water Association, and the American Association for the Advancement of Science.

Behzad Imanian is currently leading the high-throughput sequencing (HTS) Initiative in the Institute for Food Safety and Health, and he is a research assistant professor in the department of Food Science and Nutrition in Illinois Tech. His research interests include gene and genome evolution, gene transfers (HGT & EGT), organelle genome, transcriptome, proteome and metabolism, reductive evolution, tree of life, parasitology, symbioses, pathogenicity, food safety and human health.

Received: March 21, 2024. **Revised:** May 28, 2024. **Accepted:** June 17, 2024

© The Author(s) 2024. Published by Oxford University Press.

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

For commercial re-use, please contact journals.permissions@oup.com

pose a threat to animal and human health in more than one way. AMR, multidrug resistance (MDR), and the rise of “superbugs” are all related and considered as a dire global threat to public health [11, 12]. At the same time, plasmids have become a valuable tool in molecular cloning, genetic engineering, gene therapy, and drug production, all of which are being explored in order to improve our health and environment [13]. Plasmid-borne genes, especially the gene clusters for secondary metabolites, are now routinely used for the development of natural products and new drugs [14] including new antibiotics [15]. In addition, plasmids harbor genes that can be utilized in environmental conservation, for example, in bioremediation [16]. Given the ubiquity of plasmid-harboring bacteria in the environment including food, water, soil, and even air, investigating plasmids is increasingly important to human health and environmental conservation.

Studying plasmids have recently benefited greatly from the high-throughput sequencing (HTS) of the bacterial whole genomes and microbial metagenomics. By using HTS, we have learned a great deal more about plasmids, the extent and nature of their threats to human health, and their many practical uses. However, using HTS in plasmid studies is not free of challenges. For example, it is difficult to discern plasmid sequences from those of the chromosomes in the big datasets that are produced by all the existing HTS sequencers. Even after assembling the raw reads in larger contigs, the challenge remains because these assemblies usually contain many plasmid-sized contigs that have a chromosomal origin. It is, thus, crucial to develop reliable tools to distinguish plasmid sequences from chromosomal sequences in the pool of millions of reads that are resulted from sequencing of an axenic or an environmental sample.

Several methods have exploited the discernable sequence features and gene contents in plasmids versus chromosomes to develop such tools. In recent years, machine learning (ML) has been added to the list of these methods, resulting in the improvement of the new plasmid identification tools. ML algorithms take a subset of the data called the ‘training data’ as an input and examines and evaluates the correlation between ‘feature variables’ and ‘target variables’, a process called ‘learning’, in order to predict the ‘target variables’ based on the features of the new data. These tools differ in the features they exploit and/or the algorithm they use for modeling, and some perform better than others. For example, PlasFlow [17] uses the sequence signatures of *k*-mer (3–7 nt) frequency in the assembled contigs as the main feature to predict plasmids. With a test dataset of contigs of 1–1570 kbp, it achieves an accuracy of 89.5%. Deeplasmid [18] uses the features from both sequence signatures and gene content, including GC content, homopolymer, plasmid replication origin, coding density, contig length, hit to plasmid proteins, and hit to a curated Pfam database. Using a test dataset with contigs of 1–330 kbp, Deeplasmid achieves an accuracy of 84.2% and an area under the receiver operating characteristic (ROC) curve (AUC) of 89.9%. The Deeplasmid’s precision is high (up to 94.5%), but its recall is relatively low (75.6%). This means that a high percentage of the true plasmids (24.4%) are dismissed and not detected at all, and from all the predicted plasmids, many (5.5%) are predicted incorrectly. Of the remaining tools, PlasClass [19] uses logistic regression (LR) classifiers and a *k*-mer-based (3–7 nt) feature vector for fragments; PlasmidVerify [20] employs NB classifier and the gene content of cyclocontigs; and PlasForest [21] uses a homology-based random forest and a few other sequence features. These tools have reported higher or lower accuracies and speeds using their own test datasets.

Despite the availability of these tools for plasmid detection, several issues were needed to be addressed. Firstly, the overall accuracies, recalls and precisions of these tools were not satisfactory enough to identify plasmid sequences in the assembly files with high confidence. Secondly, the accuracies of the previous tools were not made comparable using the same test dataset. Thirdly, the running time required for some of these tools was too long (up to hours).

Here, we present a new plasmid identification tool, PlasmidHunter, that uses only gene content profile features to predict plasmid sequences with no reliance on the raw sequence data, sequence topology, and coverage or assembly graph. Thus, the input data for PlasmidHunter are simply any assembled sequence file produced by any modern high-throughput sequencer and assembled by any algorithm. Using the same dataset, we also demonstrate that PlasmidHunter achieves both higher accuracy (96.7%) and recall (95.1%) with reasonable speed (<8 min) in comparison to the previous tools. We also present a benchmarking of all the top tools using the same test datasets of contigs with different lengths.

Results

Database construction

In order to build a database for gene content profiling, 49 million unique protein sequences of 37 098 complete prokaryotic genomes from the National Center for Biotechnology Information (NCBI) RefSeq database were downloaded and processed. After clustering using MMseqs2 with the identity cutoff of 40% and singleton removal, 3.9 million proteins (representing 43 million proteins) were acquired and used to populate a database for gene content profiling used in the subsequent modeling and prediction steps (Fig. 1A).

Gene content profile was a better discriminator of chromosome and plasmid sequences than *k*-mer frequency profile

In order to compare the discriminative power of gene content profile and *k*-mer frequency profile features of our model, we performed principal component analysis (PCA) on representative species from all the 35 phyla (Fig. 2). The PCA results for the *k*-mer frequency profile (4–6 bp) showed that the contigs originating from plasmids and chromosomes of different species were scattered with occasional intersections or overlaps in different regions, making it difficult to visually distinguish plasmid contigs from chromosomal contigs (Fig. 2A–C). In contrast, in the PCA results for the gene content profile, all the chromosomal contigs were well separated from all the major plasmid contigs, with small amount of plasmid contigs close by. The chromosomal contigs of all the species were highly concentrated in a single region (Fig. 2D and E). These results clearly demonstrated the higher discriminative power of gene content profile in discerning plasmid and chromosomal sequences than that of the *k*-mer frequency profile. Thus, we chose gene content as the feature for our modeling.

The gene content profile PCA analysis had a significantly lower explained variance ratio for PC1 and PC2, in comparison to the *k*-mer, which can be attributed to the high dimensionality of the feature space. With 43 000 features, the gene content profiles had a much larger number of dimensions compared to the *k*-mer profiles, which had at most 5000 features. In high-dimensional

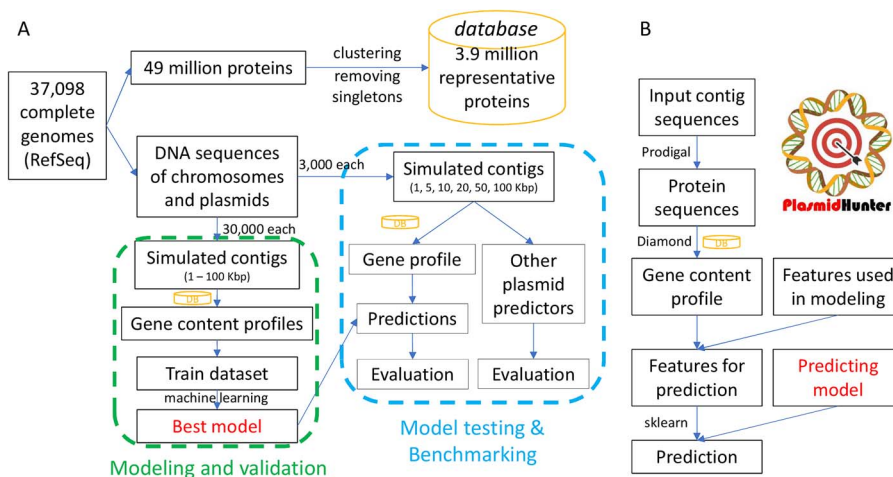


Figure 1. The workflow of the modeling, benchmarking, and pipeline construction. (A) The workflow of the database construction, modeling and validation, and the model testing and benchmarking of this study. Briefly, protein sequences of 37 098 complete genomes from NCBI RefSeq database were used and 3.9 million representative protein sequences were indexed as a database. From the complete genomes, 30 000 chromosome and 30 000 plasmid sequences were used for modeling and validation, respectively, and the remaining sequences were held back for an unbiased model testing and benchmarking. (B) The workflow of the pipeline PlasmidHunter. Input DNA sequences are first used to predict coding sequences of genes for each contig. The translated protein sequences are then used for Diamond alignment using the customized database. The gene content profile is filtered to retain only the gene features used in the modeling. The predicting model is then used to predict the chromosomal or plasmid origin of the contigs with the Python package sklearn.

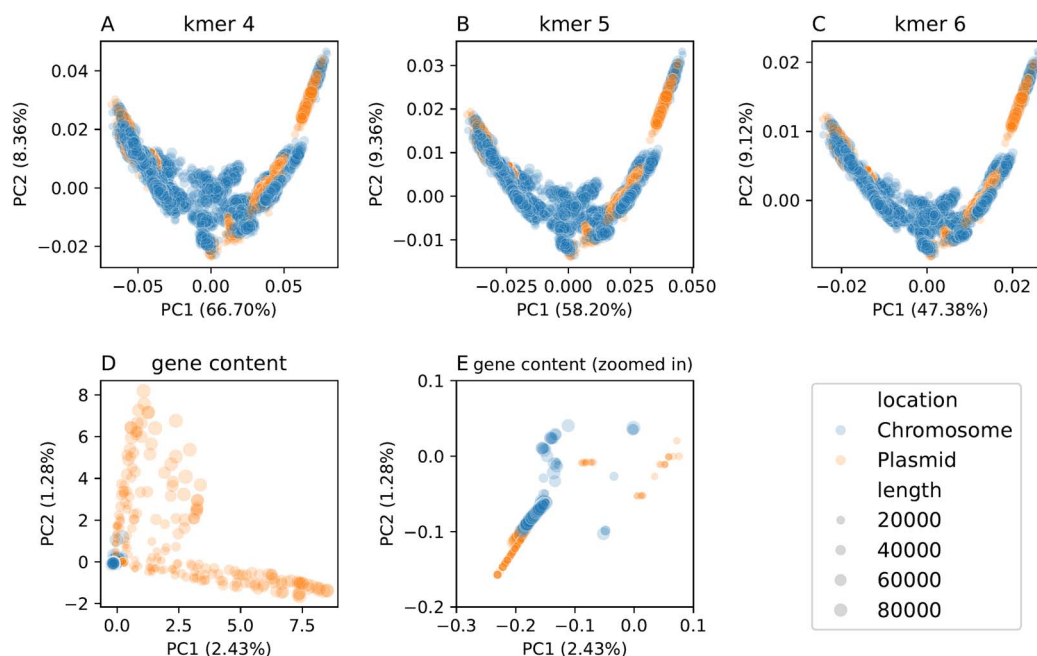


Figure 2. Comparisons between the discriminative power of kmer frequency profile and gene content profile in distinguishing plasmid and chromosome contigs. Representative species of all the phyla were used for the comparisons. For each species, 100 simulated contigs of the plasmid and chromosomal origin, respectively, were used for the PCA analysis based on the kmer 4 (A), kmer 5 (B), and kmer 6 (C) frequency profile and (D) the gene content profiles. The concentrated area of chromosomes in (D) was zoomed in (E) to better visualize the extent of separation of all the chromosome contigs and the nearest plasmid contigs. The explained variance ratios were labeled for each axis.

spaces, the variability of the data tends to be spread across multiple dimensions, resulting in lower explained variance ratios for individual principal components. The improved separation with lower explained variance ratio suggested that the gene content profiles carry more useful information for classification compared to the *k*-mer profiles.

Training dataset based on gene content profile

The downloaded 37 098 complete genomes from NCBI RefSeq database contained a total of 37 372 assigned chromosomes (some

genomes had multiple chromosomes) and 41 608 assigned plasmids. Of these sequences, 918 chromosomes and 802 plasmids were removed from further processing due to their uncharacteristic sizes (Supplementary Table 1). In total, 30 000 chromosomal and 30 000 plasmid sequences were randomly selected, and then a contig with varying length from 1 to 100 kbp (representing typical contig lengths in a genomic assembly) was randomly trimmed from each sequence (Fig. 1). After the Prodigal gene prediction and Diamond alignment, 59 896 chromosomal and plasmid contigs were annotated using the database created in this study.

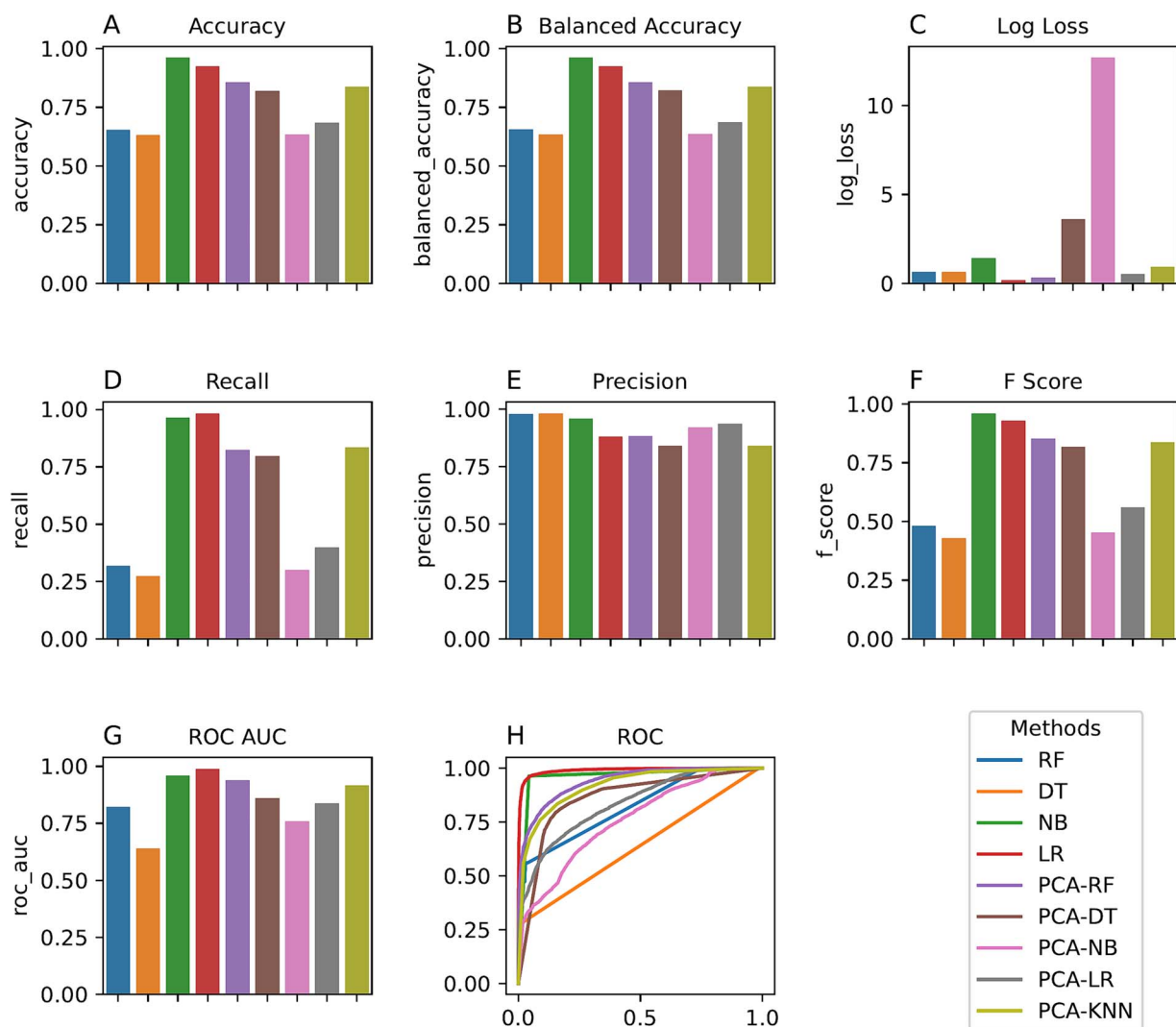


Figure 3. The performance evaluations of the models predicting the location (chromosome or plasmid) of contigs. The evaluations included (A) total accuracy, (B) balanced accuracy, (C) log loss, (D) recall, (E) precision, (F) F score, (G) AUC of ROC, and (H) ROC. Each evaluation examined the methods of RF, DT, NB, LR, and KNN. Methods beginning with PCA—refer to the modeling using PCA-transformed data.

The machine learning modeling singled out Naïve Bayes with best performance

In total, 10 models with multiple parameters were used to fit the training data (both PCA transformed and not transformed). The Naïve Bayes (NB) yielded the highest total accuracy 96.0%, and the LR achieved the second place with an accuracy of 92.4% (Fig. 3A, Supplementary Table 2). Because the validation dataset included almost an equal number of plasmid and chromosomal contigs, the balanced accuracies (Fig. 3B) were nearly the same as the total accuracies. The LR had the lowest log loss value (0.2), and thus, it provided the highest confidence on its predicting probabilities (Fig. 3C). In terms of sensitivity or Recall, the measure of true positive rate, the NB correctly predicted 96.4% of all the plasmid contigs (Fig. 3D). It also had a high precision value [true positive/(true positive + false positive)] of 95.8% (Fig. 3E), meaning that among all the predicted plasmids, 95.8% were true plasmids. Although the random forest (RF) and decision tree (DT) produced higher precision values (97.8% and 98.2%, respectively), their Recall values were very low (32.0% and 27.5%, respectively), meaning that they were too conservative in predicting a plasmid and simply ignored a big proportion of plasmids. The F score that

considers both the recall and precision were calculated for all the models, and the results indicated that the NB had the best F score, 0.96 (Fig. 3F). The indicator ROC AUC (Fig. 3G) and ROC (Fig. 3H) both indicated that the NB (AUC of 0.96) and LR (AUC of 0.99) were excellent, meaning that they could achieve a low false-positive rate (high specificity) while maintaining a high true-positive rate (high sensitivity). The k -fold cross-validation resulted in consistent accuracies of $96.4 \pm 0.2\%$. Considering all these indicators comprehensively, the NB had the best performance with high sensitivity and specificity, and it was chosen for the plasmid prediction tool development.

The PlasmidHunter outperformed other plasmid prediction tools

We generated a benchmark dataset using the genomes that were not included in the modeling (Fig. 1A, Supplementary Table 3) in order to avoid introducing any biases in the benchmarking. The benchmark data included simulated contig sequences with lengths 1, 5, 10, 20, 50, and 100 kbp, randomly selected from the genomes. We developed a pipeline named PlasmidHunter to predict sequences based on gene content using the NB model

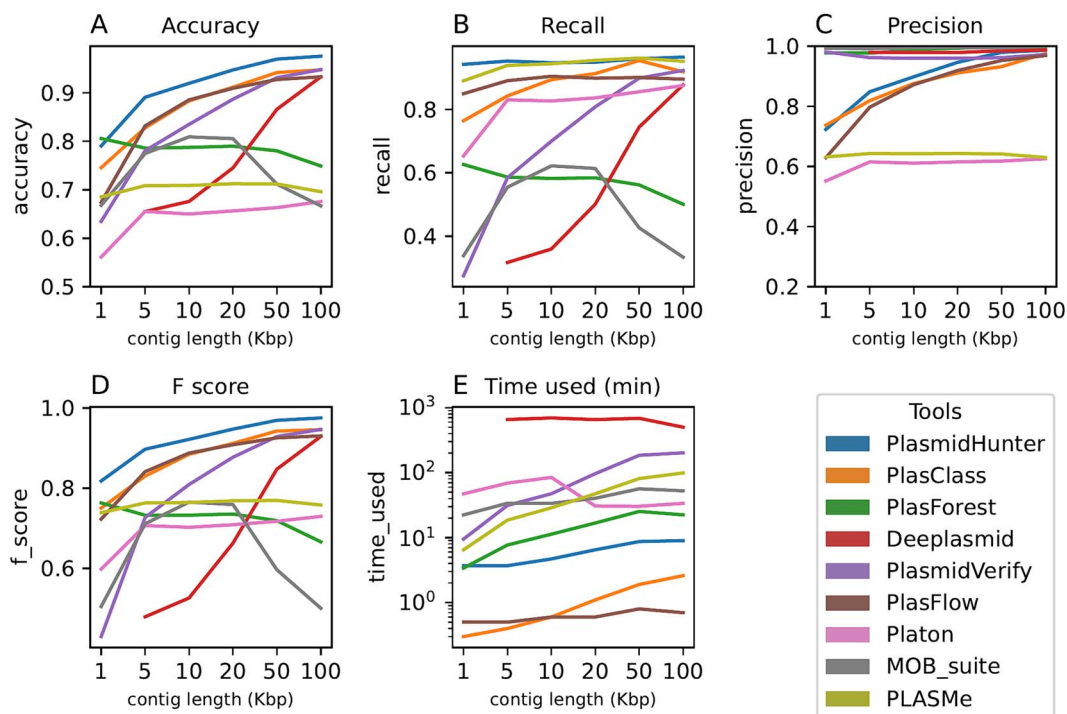


Figure 4. The performance evaluation of all the tools visualized on the benchmark data with different lengths. The benchmarking included evaluations on (A) accuracy, (B) recall, (C) precision, (D) F score, and (E) time used for running. The prediction was run on a computer with eight processors (AMD EPYC 7551, 1.2 GHz) assigned to the task, except that Deeplasmid was run on a different computer with eight processors (Intel Core i7-10510U, 1.8 GHz). The evaluation was conducted using the package scikit-learn. The log loss and ROC were not shown because some tools do not output probability of prediction.

(Fig. 1B). Using the benchmark data, we compared the accuracy and speed of PlasmidHunter with those of the eight recently developed learning-based and reference-based tools for plasmid identification: Deeplasmid, PlasClass, PlasForest, PlasmidVerify, PlasFlow, PLASMe, MOB-suite, and Platon.

With the best total accuracy, PlasmidHunter outperformed all the other tools for all the selected contig length categories except 1 kbp (5, 10, 20, 50, and 100 kb) in the six datasets with accuracies of 80.1%, 89.2%, 92.0%, 94.7%, 97.1%, and 97.6%, respectively, (Fig. 4A, Supplementary Fig. 1, Supplementary Tables 4–12). For the long contig dataset (100 kbp), PlasmidHunter achieved an accuracy of 97.6% while the accuracies of the other tools were between 66.7% and 94.8%. Overall, PlasmidHunter topped the list, and PlasClass and PlasFlow came next with similar total accuracies; they performed better than PlasmidVerify, PlasForest and Deeplasmid, PLASMe, and the two reference-based tools (Fig. 4A).

As for recall values, PlasmidHunter again outperformed all the other tools for all contig lengths (Fig. 4B). Recall represents how sensitively the tools can detect plasmid sequences. PlasmidHunter detected 94.8%–97.0% of all the plasmid sequences in the six contig datasets (Fig. 4B). PLASMe, PlasClass, Platon, and PlasFlow had close recall values and were better than PlasmidVerify, PlasForest, MOB_suite, and Deeplasmid. In terms of precision, PlasForest, Deeplasmid, PlasmidVerify, and MOB_suite were the best (between 96.0% and 100%, Fig. 4C), which means that almost all the sequences predicted as plasmids were correct. However, their sensitivity or recall values were low (Fig. 4B). Interestingly, the tools with the highest precision would generally have a lower recall, and vice versa. In terms of F score, which is the harmonic mean of recall and precision and balances the two indicators, PlasmidHunter got the top score. PlasClass and PlasFlow (Fig. 4D) came next, with similar F scores that were higher than the rest.

In terms of speed, PlasmidHunter spent between 3.3 and 8.5 min for the six benchmark datasets. PlasFlow and PlasClass were the fastest tools, completing the task in 0.3–2.6 min. Deeplasmid was the slowest and took much longer, up to hundreds of minutes.

We also tested all the tools using real metagenomics data. We downloaded 24 plasmidome data and divided them into groups based on sequence lengths (1–3, 3–5, 5–7, 7–9 kbp) for the testing. The result showed that PlasmidHunter, PlasClass and PlasFlow had better accuracies for all contig lengths than the others (Supplementary Fig. 2). Among them, PlasmidHunter outperformed the rest for all lengths except for short contigs at 1–3 kbp, with an accuracy of 79.0%. It performed better for the longer contigs with accuracies of 92.8%–95.7%. Note that the accuracy here was equivalent to recall because all the contigs were plasmid sequences in the plasmidome data. The plasmidome data served as an independent dataset for comparison, as they are unlikely to have been included in the training data of any of the tools evaluated. By using this dataset, we provide an unbiased assessment of the performance of PlasmidHunter and other tools on novel, real-world data.

Discussion

In the past two decades, researchers in life sciences have been increasingly using HTS to investigate a wide range of important subject matters in taxonomy, evolutionary biology, ecology, agriculture, environmental and animal conservation, and human health to name just a few. Consequently, the sequence databases have grown astronomically, becoming a treasure trove of information for a growing number of scientists. However, the raw sequence data that usually contain millions of reads of

different lengths need to be properly checked for quality, sorted, assembled, and annotated. For example, in metagenomic data obtained through shotgun sequencing of environmental samples that usually contain many microbes, determining which reads belong to what organism is crucial for proper analyses and correct conclusions. Even whole genome sequencing of an axenic sample could produce reads that belong to different genomic compartments such as the membrane bound mitochondria, plastids, and nucleus in eukaryotes or the naked plasmids and chromosomes in prokaryotes.

In addition to their chromosomes, many prokaryotes, both archaeobacteria and eubacteria, carry one or more plasmids. Due to their ubiquity and mobility, plasmids are important vectors of non-essential but often beneficial traits such as virulence and AMR among the closely or even distantly related bacterial species. Therefore, it is critically important to develop tools that correctly determine the location/origin of the reads, plasmids, or chromosomes, in the large sequence datasets. Although the latest plasmid-identifying tools have achieved some improvements, they still suffer from certain limitations. We have assessed the accuracies, recalls, precisions, F scores, and speed of some of the most recently developed tools to identify plasmids, and here, we introduce PlasmidHunter, a new tool that overall performs better than the other tested tools, achieving higher speeds, accuracies, and recalls with contigs of various lengths.

Our evaluations of some of the most recent plasmid-identifying tools indicated that the use of multiple features might lead to lower accuracies. Using multiple features in ML modeling is sometimes necessary; however, multiple features introduce more assumptions, each of which might not be essential or correct. According to the heuristic Occam's Razor in machine learning, all things being equal, simpler models can predict better than more complex models. For example, for its predictions, Deeplasmid uses both sequence signatures and gene contents—multitude of features that include GC content, homopolymer, plasmid replication origin, coding density, contig length, hit to plasmid proteins, and hit to a curated Pfam database among others. Some of these features such as GC content and contig length may not be significantly and/or consistently different between chromosomes and plasmids along their entire lengths and among the short and long contigs originated from each in the assembly data with millions of reads. This might explain the lower accuracy of Deeplasmid compared to the tools with simpler/fewer features (Supplementary Table 5). In contrast, PlasmidHunter, PlasFlow and PlasClass all use only one type of feature, a *k*-mer profile or a gene content profile, and they all achieve higher accuracies and speeds than other tools that use multiple features (Supplementary Tables 4, 7, and 9). Our results showed that the gene content profile was more discriminative than a *k*-mer profile for differentiating chromosomal from plasmid contigs (Fig. 2). This may explain why PlasFlow and PlasClass, which use merely a *k*-mer profile feature, had lower accuracies than PlasmidHunter.

In the benchmarking, some tools reached higher precisions for all contig lengths than others. The Deeplasmid, PlasForest, PlasmidVerify, and MOB_suite had a precision of 96.0%–100% across all the benchmark datasets (Fig. 4C). This means that almost all the contigs predicted as plasmid were true plasmid contigs. However, their low sensitivities or recall values, especially for short contigs (down to 31.7%, Fig. 4B), mean that they detect a small percentage of all the plasmid contigs in the dataset. As a result, their F scores were also much lower than the others (Fig. 4D). In comparison, PlasmidHunter had a more balanced recall and precision than all the other tools.

Compared to the previous top plasmid-identifying tools and based on the benchmarking using datasets of sequences with different lengths, PlasmidHunter achieved the best total accuracies and recalls or sensitivities for all contig lengths (Fig. 4A and B). On the downside, PlasmidHunter showed a lower precision than some of the other tools for short contigs (72.0% and 84.8% for 1- and 5-kbp contigs, respectively).

In certain applications, where the goal is to identify true plasmids, precision—the accuracy of predicted plasmid contigs—is prioritized, even if only a few true plasmids are identified. Our tool, while not leading in precision, performs well, with a precision rate of >85% for contigs larger than 5 kbp, and this rate increases notably to 90% for contigs larger than 10 kbp, especially when users focus on longer contigs. On the other hand, there are applications with the objective of identifying as many candidate plasmids as possible prior to validation or screening, for which a high-recall tool is necessary to predict a greater number of potential plasmids. PlasmidHunter outperformed all the others in recall. Ideally, a tool offering a balanced recall and precision, represented by the F score, is most suited for practical applications. Among competing tools, PlasmidHunter boasts the highest F score. Based on the benchmarking results of both simulated contigs and real metagenomics data, we recommend using PlasmidHunter over the other tools to predict contigs longer than 3 kbp to achieve good recall and precision.

Conclusion

Here, we have made rigorous comparisons between recently published plasmid-prediction tools using independent benchmark datasets of contigs with different lengths. We showed that the use of multiple and complex features does not necessarily result in higher accuracy in modeling. Our study also provides useful insights into feature selection as well as ML algorithms for modeling in sequence classifications. Finally, we present our tool, PlasmidHunter, achieving the highest accuracy.

Materials and methods

Database preparation

The NCBI prokaryotic genome databases were searched for with the filtration assembly level of 'complete' (<https://www.ncbi.nlm.nih.gov/genome/browse#/prokaryotes/>), and the results were downloaded as metadata. As of 6 November 2023, there were 37 098 complete genomes of prokaryotes. A custom Python script was used to download all the protein sequences of the genomes using the URLs in the metadata with multiple processes. The protein sequences were then clustered using MMseqs2 (version 15.6f452) [22] with sequence identity cutoff of 0.4, alignment coverage of 0.8 and `-cov-mode` of 0 (bidirectional). The clusters with <2 sequences were removed. The representative sequences of the remaining 3.9 million clusters (i.e. one representative sequence from each cluster) were used as a database for gene content profiling.

Gene content annotation

Gene content profiling was conducted by alignment to the sequences in the protein database. First, genes were predicted using Prodigal (version 2.6.3) [23] in the meta or anonymous mode, allowing genes to run off edges. The protein sequences were then used to align against the database using BLASTp mode of DIAMOND (version 2.0.15.153) [24]. Query coverage of 80%, protein sequence identity of 30%, and E value of 1e-5 were used to

filter the hits (`--max-target-seqs 1 --max-hsps 1 --evaluate 1e-5 --id 30 --query-cover 70`). The protein IDs of the hits with the highest alignment scores were assigned to the queries. The gene content profile was then summarized into a data frame in Python.

Comparing the discriminatory power of gene content-based and k-mer-based methods

The metadata table from the NCBI was used to select bacterial genome for the comparison. There were 37 098 genomes belonging to 35 phyla, and one representative genome of each phylum was selected. Genome sequences were downloaded using `ncbi-genome-download` (version 0.3.0) and 100 simulated contigs of chromosomes and plasmids, respectively, with random lengths between 5 and 100 kbp generated for each genome. The gene content profile of the simulated contigs was generated as described above. The k-mer frequency profile ($k=4, 5, \text{ and } 6$, respectively) of the simulated contigs was generated using custom codes with Python package `collections.Counter` and `Pandas`. The gene content and k-mer frequency profiles were then compared by dimensionality reduction with principal component analysis (PCA). The Python package `sklearn.decomposition.PCA` was used to conduct the PCA analysis with two components. The outputs were then visualized with a Python package `seaborn` (version 0.11.2).

Training data set for modeling

The gff files of the 37 098 complete prokaryotic genomes from NCBI RefSeq database were downloaded using the URLs in the metadata and parsed using the Python package `BCBio` in a multi-processing mode. For each gff record, the sequence length, location (plasmid or chromosome) and a list of protein IDs in the original order were extracted and saved in a json file. To exclude any false annotation of sequence location, the sequence was defined as chromosomal if the annotation was “chromosome” and the size was >900 kbp, and as plasmid if the annotation was “plasmid” and the size was <600 kbp. The plasmid sequences <1 kbp were removed.

For the training dataset, 30 000 plasmid and 30 000 chromosome sequences were randomly selected and 3000 of each from the rest of the sequences were set aside for model testing and evaluation. The DNA sequences of genomes were downloaded using the URLs in the metadata. For each selected sequence, a fragment with a random start point and a random length (from 1 to 100 kbp, representing typical contig lengths in a genomic assembly) was selected. If a sequence length was less than the random length, the whole sequence would be used. The simulated contigs would then be annotated as mentioned above and a feature table of gene presence and absence of all the simulated contigs would be generated in Python `Pandas`, using the representative protein IDs as features.

Machine learning modeling

The Python package `scikit-learn` (version 1.0.2) was used for the modeling. The training data were first split into training dataset (75%) and validation dataset (25%) using the function `train_test_split`. For the modeling, the following algorithms were tested to fit the models: Decision Tree (with maximum depth of 5, 10, 15, and 20); Random Forest (with maximum depth of 5, 10, 15, and 20); NB; LR; Support Vector Machine (with regularization 0.1, 1, and 10); and K Nearest Neighbors ($n=7$). The validation dataset was then used to calculate the accuracy of the classifications. Meanwhile, the training data were transformed with PCA transformation (with PC number 30), and the transformed data were subjected to the same processes as mentioned above.

Model validation

The validation dataset (25% of training data) was held back to be used for the model validation. For each model, the prediction scores, hence the probabilities of each class, were calculated using the function `predict_proba`. A custom function accepting inputs of true classification (`y_true`) and prediction scores (`y_score`) was used to evaluate each model’s performance with the package `sklearn.metrics`. The total accuracy was calculated with the function `accuracy_score`. The Log Loss was calculated with the function `log_loss`. The recall (true-positive rate) and precision [$\text{true positive}/(\text{true positive} + \text{false positive})$] were calculated with the function `recall_score` and `precision_score`, respectively. An F score was based on the harmonic mean of the recall and precision. Finally, an area under curve (AUC) of the receiver operating characteristic (ROC) curve was calculated with the function `roc_auc`. The best model was “dumped” as a local binary file using `pickle`. To thoroughly assess the model’s accuracy, we implemented a k-fold cross-validation approach. This involved utilizing the `KFold` class from the `sklearn.model_selection` package. Specifically, we divided the training data into 10 distinct subsets, ensuring a comprehensive evaluation by systematically training and validating the model across each subset.

Benchmark data for model test and evaluation

As mentioned above, 30 000 chromosome and plasmid sequences, respectively, from the 37 098 complete prokaryotic genomes from RefSeq database, were used for the modeling and validation. The remaining data were held back from the modeling for testing and unbiased evaluation. The balanced datasets including an equal number of chromosome and plasmid sequences (3000 each) were randomly selected. Contigs of different lengths, including 1, 5, 10, 20, 50, and 100 kbp, were randomly simulated from each sequence and were used for the gene content profiling and prediction. The contigs were annotated using `DIAMOND` search against the database described above. The gene content profile was then processed to ensure that it had the same features that were used/included in the modeling data. The model was loaded using `pickle` and was used to predict the feature data. The performance of the model on the data of different sequence lengths was then evaluated with the package `sklearn.metrics` as mentioned in the model validation.

In addition, to test and compare the tools with real metagenomic data, real plasmidome data were downloaded from a global sewage plasmidome project [25] from the European Bioinformatics Institute (EBI) database with the accession ERZ1694234 to ERZ1694257. According to the study, the chromosomal DNA was degraded using `Plasmid-Safe ATP-dependent DNase`, and the plasmid DNA was amplified through rolling-circle amplification using `phi29 DNA polymerase`. Further sequence analysis has shown that 96% of the assembled contigs (1.0–17.4 kbp, average length 1.9 kbp) were circular, concluding that they were originated from plasmids. We divided the sequences into multiple groups based on sequence lengths (1–3 kbp, 3–5 kbp, 5–7 kbp, 7–9 kbp). As a result, there were 211 832, 20 103, 4577, and 395 sequences, respectively, from the four groups. We randomly selected at most 1000 sequences from each groups and used them to test all the tools for comparison.

Construction of the gene content-based plasmid prediction pipeline

A pipeline named `PlasmidHunter` was developed using the validated model to predict plasmid sequences from input contigs. First, the input contigs are filtered to remove short sequences

(<1 kbp). Prodigal (version 2.6.3) is used to predict gene and protein sequences. Diamond (version 2.0.15.153) is used to search the protein sequences against the custom database. A gene content profile is generated using the alignment results based on the features used in the modeling step. The gene content profile is then used as features for prediction using the Python package sklearn (version 1.0.2) and the NB model.

Benchmarking of multiple plasmid predictors

Nine tools, namely, PlasmidHunter, PlasClass [19], PlasFlow [17], PlasmidVerify [20], PlasForest [21], Deeplasmid [18], Platon [26], MOB_suite [27], and PLASMe [28] were tested using the benchmark datasets with different contig lengths. All the tools were run following the manuals provided, on a high-performance computer (HPC) with eight processors (AMD EPYC 7551, 1.2 GHz) assigned to the tasks, except Deeplasmid, which was run on a different computer with eight processors (Intel Core i7-10510U, 1.8 GHz) because the Deeplasmid cannot be limited to use only eight processors through the command line. The running was timed using the Python package, time. The outputs were parsed to retrieve the prediction of each contig and the corresponding probability if any. If the prediction output only included classes without probabilities, 0 for chromosomes and 1 for plasmids were used as the probabilities of plasmid. For the unpredicted samples, half of them were assigned as plasmids and half as chromosomes for a fair comparison and evaluation. The Python package sklearn.metrics was used to evaluate the prediction results in terms of total accuracy, balanced accuracy, recall, precision, F score, and AUC ROC. The Python package seaborn was used to visualize the evaluation results.

Key Points

- Plasmids are transmissible DNA segments that play a crucial role in bacterial evolution and ecology.
- Plasmids carry beneficial genes for hosts, contributing to resistance, virulence, toxicity, and pathogenicity.
- HTS helps understand bacterial evolution and ecology, but discerning plasmid from chromosomal sequences remains challenging.
- PlasmidHunter, a new machine learning tool, predicts plasmid sequences using gene content profile features.
- PlasmidHunter outperforms existing tools in accuracy and recall; a benchmarking of top tools is presented.

Acknowledgments

The authors acknowledge the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) for the support.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This publication is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) (Grant No. 5U19FD005322) as part of an award totaling

\$3 856 000 with 0% financed with nongovernmental sources. The funding body did not play a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official views of, nor endorsement by, the FDA, HHS, U.S. Government, or Illinois Institute of Technology. For more information, please visit <https://www.fda.gov/>.

Availability of the scripts and the benchmark data sets of this study

All the scripts for data processing and analysis and the benchmark datasets with different contig lengths have been deposited on GitHub (<https://github.com/tianrenmaogithub/PlasmidHunter>).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The source code of PlasmidHunter is available on GitHub and FigShare for plasmid prediction analysis (<https://github.com/tianrenmaogithub/PlasmidHunter>, <https://doi.org/10.6084/m9.figshare.25106219.v1>). The scripts of this project and the benchmark data are available on Zenodo (<https://zenodo.org/records/10433596>).

Author contributions

B.I.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. R.T.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. J.Z.: Formal analysis, Writing – review & editing. All authors have read and agreed to the published version of the manuscript.

References

1. Kothari A, Wu Y-W, Chandonia J-M, et al. Large circular plasmids from groundwater Plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. *MBio* 2019;**10**:e02899–18. <https://doi.org/10.1128/mBio.02899-18>.
2. Rozhon W, Petutschnig E, Khan M, et al. Frequency and diversity of small cryptic plasmids in the genus *Rahnella*. *BMC Microbiol* 2010;**10**:56. <https://doi.org/10.1186/1471-2180-10-56>.
3. Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology* 2015;**6**:242. <https://doi.org/https://www.frontiersin.org/articles/10.3389/fmicb.2015.00242>.
4. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, et al. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat Rev Microbiol* 2021;**19**:347–59. <https://doi.org/10.1038/s41579-020-00497-1>.

5. Silver S, Misra TK. Plasmid-mediated heavy metal resistances. *Annu Rev Microbiol* 1988;**42**:717–43. <https://doi.org/10.1146/annurev.mi.42.100188.003441>.
6. Martínez-Martínez L, Pascual A, Jacoby GA. Quinolone resistance from a transferable plasmid. *Lancet* 1998;**351**:797–9. [https://doi.org/10.1016/S0140-6736\(97\)07322-4](https://doi.org/10.1016/S0140-6736(97)07322-4).
7. Tran JH, Jacoby GA. Mechanism of plasmid-mediated quinolone resistance. *Proc Natl Acad Sci U S A* 2002;**99**:5638–42. <https://doi.org/10.1073/pnas.082092899>.
8. Meng M, Li Y, Yao H. Plasmid-mediated transfer of antibiotic resistance genes in soil. *Antibiotics (Basel)* 2022;**11**:525. <https://doi.org/10.3390/antibiotics11040525>.
9. Rice LB. Mechanisms of resistance and clinical relevance of resistance to β -lactams, Glycopeptides, and fluoroquinolones. *Mayo Clin Proc* 2012;**87**:198–208. <https://doi.org/10.1016/j.mayocp.2011.12.003>.
10. Krause KM, Serio AW, Kane TR, et al. Aminoglycosides: an overview. *Cold Spring Harb Perspect Med* 2016;**6**:a027029. <https://doi.org/10.1101/cshperspect.a027029>.
11. Larsson DGJ, Flach C-F. Antibiotic resistance in the environment. *Nat Rev Microbiol* 2022;**20**:257–69. <https://doi.org/10.1038/s41579-021-00649-x>.
12. CDC. CDC's response to a global emerging threat [internet]. Centers for Disease Control and Prevention. 2022. Available from: <https://www.cdc.gov/drugresistance/solutions-initiative/stories/ar-global-threat.html>.
13. Doghaither HA, Gull M, Doghaither HA, Gull M. Plasmids as genetic tools and their applications in ecology and evolution [internet]. *Plasmid IntechOpen* 2019. Available from: <https://doi.org/10.5772/intechopen.85705>. <https://www.intechopen.com/state.item.id>.
14. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod* 2020;**83**:770–803. <https://doi.org/10.1021/acs.jnatprod.9b01285>.
15. Schneider YK. Bacterial natural product drug discovery for new antibiotics: strategies for tackling the problem of antibiotic resistance by efficient bioprospecting. *Antibiotics (Basel)* 2021;**10**:842. <https://doi.org/10.3390/antibiotics10070842>.
16. Suenaga H, Koyama Y, Miyakoshi M, et al. Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. *ISME J* 2009;**3**:1335–48. <https://doi.org/10.1038/ismej.2009.76>.
17. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;**46**:e35. <https://doi.org/10.1093/nar/gkx1321>.
18. Andreopoulos WB, Geller AM, Lucke M, et al. Deepplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Res* 2022;**50**:e17. <https://doi.org/10.1093/nar/gkab1115>.
19. Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. *PLoS Comput Biol* 2020;**16**:e1007781. <https://doi.org/10.1371/journal.pcbi.1007781>.
20. Antipov D, Raiko M, Lapidus A, et al. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res* 2019;**29**:961–8. <https://doi.org/10.1101/gr.241299.118>.
21. Pradier L, Tissot T, Fiston-Lavier A-S, et al. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinformatics* 2021;**22**:349. <https://doi.org/10.1186/s12859-021-04270-w>.
22. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. <https://doi.org/10.1038/nbt.3988>.
23. Hyatt D, Chen G-L, LoCascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119. <https://doi.org/10.1186/1471-2105-11-119>.
24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60. <https://doi.org/10.1038/nmeth.3176>.
25. Kirstahler P, Teudt F, Otani S, et al. A peek into the Plasmidome of global sewage. *mSystems* 2021;**6**:e0028321. <https://doi.org/10.1128/mSystems.00283-21>.
26. Schwengers O, Barth P, Falgenhauer L, et al. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* 2020;**6**:mgen000398. <https://doi.org/10.1099/mgen.0.000398>.
27. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;**4**:e000206. <https://doi.org/10.1099/mgen.0.000206>.
28. Tang X, Shang J, Ji Y, et al. PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer. *Nucleic Acids Res* 2023;**51**:e83. <https://doi.org/10.1093/nar/gkad578>.