MICROBIAL GENOMICS

METHOD

van der Graaf-van Bloois *et al.*, *Microbial Genomics* 2021;7:000683

DOI 10.1099/mgen.0.000683

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning

Linda van der Graaf-van Bloois[1,2], Jaap A. Wagenaar[1,2,3] and Aldert L. Zomer[1,2,*]

Abstract

Antimicrobial-resistance (AMR) genes in bacteria are often carried on plasmids and these plasmids can transfer AMR genes between bacteria. For molecular epidemiology purposes and risk assessment, it is important to know whether the genes are located on highly transferable plasmids or in the more stable chromosomes. However, draft whole-genome sequences are fragmented, making it difficult to discriminate plasmid and chromosomal contigs. Current methods that predict plasmid sequences from draft genome sequences rely on single features, like *k*-mer composition, circularity of the DNA molecule, copy number or sequence identity to plasmid replication genes, all of which have their drawbacks, especially when faced with large single-copy plasmids, which often carry resistance genes. With our newly developed prediction tool RFPlasmid, we use a combination of multiple features, including *k*-mer composition and databases with plasmid and chromosomal marker proteins, to predict whether the likely source of a contig is plasmid or chromosomal. The tool RFPlasmid supports models for 17 different bacterial taxa, including *Campylobacter*, *Escherichia coli* and *Salmonella*, and has a taxon agnostic model for metagenomic assemblies or unsupported organisms. RFPlasmid is available both as a standalone tool and via a web interface.

DATA SUMMARY

- RFPlasmid is a Linux-based tool and the software is available at https://github.com/aldertzomer/RFPlasmid.
- A pip package is available for installation of RFPlasmid.
- A platform-independent web interface for RFPlasmid is available at http://klif.uu.nl/rfplasmid/.
- RFPlasmid databases containing all plasmid proteins are available at http://klif.uu.nl/download/plasmid_db/.
- Training data sets are available at http://klif.uu.nl/download/plasmid_db/trainingsets2/trainingsfiles_zip.
- All databases and files are available on Zenodo (https://doi.org/10.5281/zenodo.3968422).
- Supporting data can be found on Figshare: https://figshare.com/s/b82569f2d5cd02b099cc

INTRODUCTION

Many bacterial species carry plasmids, extrachromosomal mobile genetic elements that can transfer from one bacterium to another [1]. They often replicate autonomously in the host using a variety of replication systems. Generally, they are circular; however, some species carry linear plasmids [2, 3]. Plasmids often carry genes that provide a benefit to the host, such as additional metabolic capabilities [4], antimicrobial-resistance (AMR) genes [5] and virulence factors that affect host invasion and infection, including type IV secretion systems, toxins, adhesins, invasins and antiphagocytic factors [6, 7].

Conjugative transfer of plasmids is considered the most important way of spreading AMR among bacteria [8]. There is a growing concern about the possibility of AMR transmission via the food chain [9]. Furthermore, the integration of AMR genes in chromosomes is a worrying development for new epidemic strains, as it provides a mechanism for vertical transmission of AMR genes without the potential fitness deficit associated with the maintenance of plasmids [10]. For molecular epidemiology purposes and risk assessment, the identification of chromosomal and plasmid sequences provides fundamental knowledge regarding the transmission

Received 28 June 2021; Accepted 07 September 2021; Published 30 November 2021

Author affiliations: [1]Faculty of Veterinary Medicine, Department of Infectious Diseases and Immunology, Utrecht University, Utrecht, The Netherlands; [2]WHO Collaborating Centre for Reference and Research on Campylobacter and Antimicrobial Resistance from an One Health Perspective/OIE Reference Laboratory for Campylobacteriosis, Utrecht, The Netherlands; [3]Wageningen Bioveterinary Research, Lelystad, The Netherlands.

***Correspondence:** Aldert L. Zomer, a.l.zomer@uu.nl

Keywords: antibiotic resistance; chromosome; machine learning; plasmid; whole-genome sequencing.

Abbreviations: AMR, antimicrobial resistance; CDS, coding sequence; NCBI, National Center for Biotechnology Information; OOB, out-of-bag.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables and two supplementary figures are available with the online version of this article.

of AMR and is essential in surveillance of bacteria with plasmid-associated AMR. Molecular identification of plasmid and chromosomal genotypes can distinguish whether the spread of AMR genes is driven by epidemic plasmids to different hosts or by clonal spread of bacterial organisms.

Many molecular epidemiology studies using short-read Illumina sequences are available for resistant organisms and the number of sequenced genomes available is in the hundreds of thousands [11–13]. These existing datasets could provide a wealth of information on plasmid dissemination, were it not for one major drawback: assembly of short-read sequencing data results in hundreds of contigs that must be individually characterized as to their origin from either plasmid or chromosomal DNA.

Multiple bioinformatic methods have been described to predict plasmids *in silico*, e.g. cBar by using distinguishing pentamer frequencies [14], PlasmidSPAdes by using assembly coverage [15], Recycler [16], PlasmidFinder by using replicon sequences [17], PLACNET by combining assembly information, comparison to reference sequences and plasmid-diagnostic sequence features [18], PLAScope by using chromosomal and plasmid databases [19], MLPlasmids by using pentamers and machine learning [20], and Platon by using replicon distribution differences of protein-encoding genes and contig circularity [21]. The predictions with some methods suffer from a low sensitivity or specificity [22], or are optimized for one specific bacterial genus and cannot be used for metagenomics.

In this study, we present our tool RFPlasmid, a novel approach for the prediction of bacterial plasmid sequences in contigs from short-read assemblies, with models for 17 different bacterial genera and a taxon agnostic model. We compared RFPlasmid to other available tools and show it that it performs equally well or better when using taxon-specific models. We identified genomic signatures of plasmid and chromosomal sequences based on 5 bp *k*-mers, a custom plasmid protein database with >193000 entries, a database of known replicons [23], single-copy chromosomal marker genes [24], contig lengths and gene counts. We trained a Random Forest model on more than 8000 pseudo assemblies from bacterial chromosomes and plasmids, and validated our approach using both the out-of-bag (OOB) error rate of Random Forest, and an independently generated dataset of plasmid and genomic contigs. Our prediction model is optimized for genome assemblies of 17 different genera and metagenomics, outperforming any other tool currently available. Additionally, we have identified potential factors responsible for prediction errors and propose downstream analyses to alleviate these problems.

Impact Statement

Antimicrobial-resistance (AMR) genes in bacteria can rapidly spread when the genes are located on plasmids. For molecular epidemiology purposes and risk assessment, it is important to know whether an AMR gene is located on highly transferable plasmids or on the more stable chromosomes. Whole-genome sequencing makes it easy to determine whether a strain contains a resistance gene. However, it is not easy to determine whether the gene is chromosomal or plasmid located, since classification of plasmid and chromosomal contigs is difficult. RFPlasmid is able to predict whether the likely source of short-read assembly contigs are chromosomal or plasmid. The tool is optimized for 17 different bacterial taxa, including *Campylobacter*, *Escherichia coli* and *Salmonella*, and can also be used for metagenomic assemblies.

THEORY AND IMPLEMENTATION

Implementation

RFPlasmid extracts feature information from whole-genome sequence contigs, and by using a Random Forest model, the likely source (plasmid or chromosomal) of the contigs is predicted. The tool supports 17 different bacterial species or taxa, including *Bacillus*, *Borrelia*, *Burkholderia*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Cyanothece*, *Enterobacteriaceae*, *Enterococcus*, *Lactobacillus*, *Lactococcus*, *Listeria*, *Pseudomonas*, *Rhizobium*, *Staphylococcus*, *Streptomyces* and *Vibrio*, and a taxon agnostic model for unknown unsupported organisms or for metagenomics data. This taxon agnostic model is called 'generic'. A flow scheme describing the procedure is given in Fig. 1. Furthermore, the tool contains an easy-to-use training option with which additional models can be added.

Input

Contigs from short-read assemblies in FASTA format are used as input files. The web interface takes a single genome, the command line tool can process up to several thousand genomes from a single folder.

Single-copy chromosomal marker genes

CheckM [24] predicts ORFs of the contigs using Prodigal [25] and determines whether these encode taxa specific single-copy marker genes. The number of specific marker genes per contig is counted and saved.

Plasmid marker proteins

Two different reference databases with plasmid maker proteins are used: the plasmid replicon database and the plasmid protein database. The plasmid replicon database consists of known plasmid replication proteins, downloaded from the database of PlasmidFinder [23] (accession date 22 May 2017). The plasmid protein database was generated with plasmid proteins from all bacterial taxa from the National Center for Biotechnology Information (NCBI) GenBank (accession date 22 May 2017) and the plasmid database of the MOB-suite v1.4.1 [26]. Near-identical proteins were clustered using USearch v5.2.32 [27], resulting in a database with 193 176 plasmid proteins.

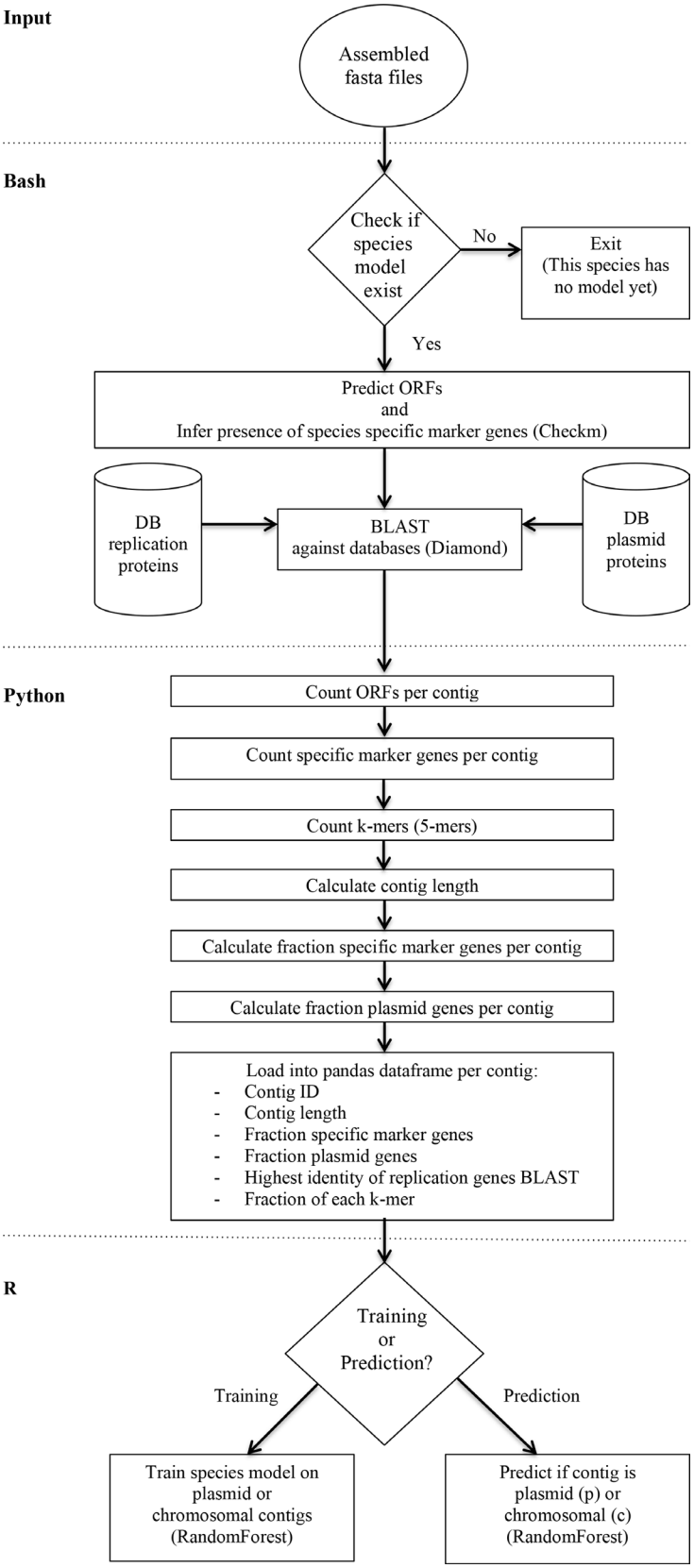


Fig. 1. Flow diagram of RFPlasmid.

RFPlasmid uses DIAMOND searches [28] against the two plasmid reference databases, BLASTX for the replicon database and BLASTP for the protein database, with default settings and an *E* value cut-off of 1e−30. For each contig, the BLASTX replicon hit with the highest identity is selected and the number of BLASTP hits with the plasmid protein database is counted.

k-mer profiles

Two different methods of *k*-mer counting are implemented: the standard method counting the number of nucleotide pentamers (5-mers) using Python (default), and the faster, optional method JELLYFISH [29]. A *k*-mer size of 5 is used, because this size outperformed 3-mers, 4-mers and 6-mers [14]. The fraction of each 5-mer is calculated in the Pandas dataframe by dividing the counted number of 5-mers by the total number of 5-mers in the contig.

Classification using Random Forest models

A Python Pandas dataframe is generated, to structure all the different features of the query whole-genome sequence contigs, including contig name, contig length, fraction-specific maker genes, fraction plasmid genes, highest replication gene identity and *k*-mer fractions. The Pandas dataframe is exported as a csv, which is imported in R for training or classification using the Random Forest library [30].

Training data sets

The training data sets were made as follows: complete and identified chromosomal and plasmid sequences were downloaded from NCBI GenBank (accession date 7 November 2017), and for *Listeria*, plasmid sequences were downloaded from NCBI GenBank with accession date 30 September 2019. Simulated reads of 500 bp each were generated with 50× coverage using the gen-single-reads script (https://github.com/merenlab/reads-for-assembly). Assembly was performed using SPAdes v3.11.1 [31] with default settings. Contigs smaller than 200 bp were removed. Table 1 shows the assemblies of the developed training data sets of each taxon. The taxon agnostic model (generic) was created by combining all chromosomal and plasmid contigs from the taxon-specific models together.

Random Forest models were trained using 5000 trees. Class imbalances were solved by making use of the sampsize option, whereby 66% of the smallest class was selected as option in sampsize for both classes when training each tree in the forest to prevent class imbalance errors and error inflation [32]. Random Forest uses an internal validation where 66% of the contigs of the training sets are used for training and 33 % are used for testing per tree in the Random Forest. The output of every tree is averaged and results in the OOB error, which is a minor overestimation of the actual error [32].

For benchmarking RFPlasmid and comparison with existing tools, RFPlasmid prediction analysis was performed using the prediction mode on the training data sets, and this prediction was compared with the output of the other tools: cBar [14], PLAScope [19], MLPlasmids [20] and Platon [21].

External validation

To investigate the performance of RFPlasmid on non-simulated data, we downloaded the Illumina and Nanopore reads of 24 multidrug-resistant *Escherichia coli* genomes from ENA (European Nucleotide Archive) from BioProjects PRJNA505407 and PRJNA387731, which were also used by Schwengers *et al*. [21]. We performed both hybrid assembly using Unicycler v0.4.9b [33] and short-read-only assembly with SPAdes (v13.3.0). We could assemble 22 isolates into distinct chromosomal and plasmid contigs using Unicycler. Isolates V232 and V92 were excluded after inspection of the sequence graphs using Bandage v0.8.1 [34], as chromosomal and plasmid contigs could not be distinguished. Contigs larger than 200 bp from the SPAdes assemblies were aligned against the corresponding complete hybrid assembly using Last v984 [35] and the best scoring hits against plasmid and chromosome contigs were collected. In total, 85 contigs (153 kb) of the 2832 (110 Mbp) contigs in the entire dataset were discarded as they had identical hits on both chromosome and plasmid.

Phage, resistance and transposase gene prediction within contigs

The presence of phage genes and resistance genes in assembled contigs of the training data were determined by performing a DIAMOND (v0.9.30) search against the ProphET phage database [36] using an *E* value cut-off of 1e−10 and the Resfinder database (accessed 01-07-2020) with a cut-off of 90 % identity and 60 % coverage (identical to the default settings of the online version of Resfinder). The presence of transposase-encoding genes was assessed by aligning encoded proteins using HMMER3 (v3.1b2) (http://hmmer.org/) against the transposase database of ISEscan [37] with an *E* value cut-off of 1e−30.

Software availability

The operating system for RFPlasmid is Linux and the software is available at https://github.com/aldertzomer/RFPlasmid. The databases containing all plasmid proteins are available at http://klif.uu.nl/download/plasmid_db/ and all training data are available at http://klif.uu.nl/download/plasmid_db/trainingsets2/trainingsfiles_zip, and all databases and files can be found on Zenodo (https://doi.org/10.5281/zenodo.3968422). A platform-independent web interface for RFPlasmid is available at http://klif.uu.nl/rfplasmid/. A pip and conda package are available for installation of RFPlasmid. The pip package instals most requirements except DIAMOND, JELLYFISH and R. CheckM requires installation of an external database.

Table 1. Assemblies of the developed training data sets

Taxon	No. of chromosomes	No. of plasmids	No. of generated contigs for training data sets (chromosome/plasmid)	Total no. of bp
Bacillus	377	291	20 055 (15 736/4319)	1.77e+09
Borrelia	28	23	1564 (110/1454)	3.32e+07
Burkholderia	211	47	26 256 (25 139/1118)	1.48e+09
Campylobacter	197	406	5423 (3652/1771)	3.42e+08
Clostridium	100	46	6537 (6044/493)	4.09e+08
Corynebacterium	166	63	4614 (4350/264)	4.31e+08
Cyanothece	5	6	634 (399/235)	2.95e+07
Enterobacteriaceae	151	2297	28 544 (13 621/14 923)	9.07e+08
Enterococcus	57	44	6270 (3693/2576)	1.73e+08
Lactobacillus	206	110	19 412 (15 610/3802)	5.17e+08
Lactococcus	37	76	3423 (2104/1319)	9.04e+07
Listeria	142	73	2685 (2371/200)	4.24e+08
Pseudomonas	254	42	18 645 (17 636/1009)	1.58e+09
Rhizobium	52	51	4241 (1573/2668)	3.50e+08
Staphylococcus	247	136	9124 (7763/1361)	6.81e+08
Streptomyces	82	64	6449 (6357/92)	7.03e+08
Vibrio	123	41	11 265 (10 282/983)	5.91e+08
Taxon agnostic model (generic)	2958	3937	222 723 (194 597/28 126)	1.19e+10

RESULTS

Classification results on training data

The number of plasmid contigs of the training data sets varied between 127 and 11 513 plasmid contigs per taxon, with the *Enterobacteriaceae* set having the highest number of plasmid contigs (Table 1). We compared the predicted contig location to the known contig location with plasmid contigs correctly classified as plasmid (called 'plasmid correct'), chromosomal contigs correctly classified as chromosomal (called 'chromosome correct'), chromosomal contigs incorrectly classified as plasmids (called 'chromosome incorrect') and plasmid contigs incorrectly classified as chromosomal (called 'plasmid incorrect'). Results are determined in percentages, both calculated as bp of each predicted contig divided by the total bp as well as percentages of correctly and incorrectly predicted contigs (Fig. 2a, b, Table S1, available with the online version of this article), where the bp percentages are the best approach to determine the prediction performance of RFPlasmid, as very small contigs with repetitive sequences make up a large part of the number of contigs, but attribute little to plasmid or chromosomal sequences.

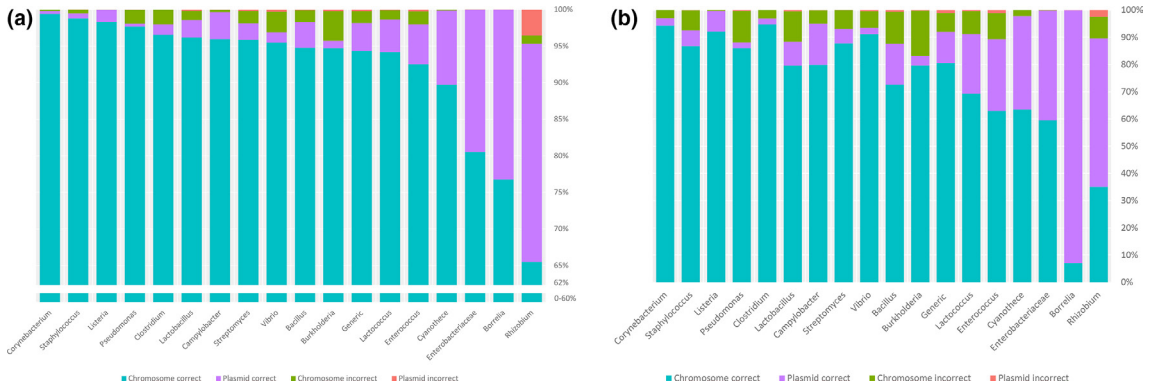


Fig. 2. Performance of RFPlasmid models on training data. Shown are the OOB performance in percentages, calculated as (a) predicted bp divided by the total number of bp and (b) predicted contigs divided by the total number of contigs, coloured as plasmid correct (purple), chromosome correct (blue), chromosome incorrect (green) and plasmid incorrect (red).

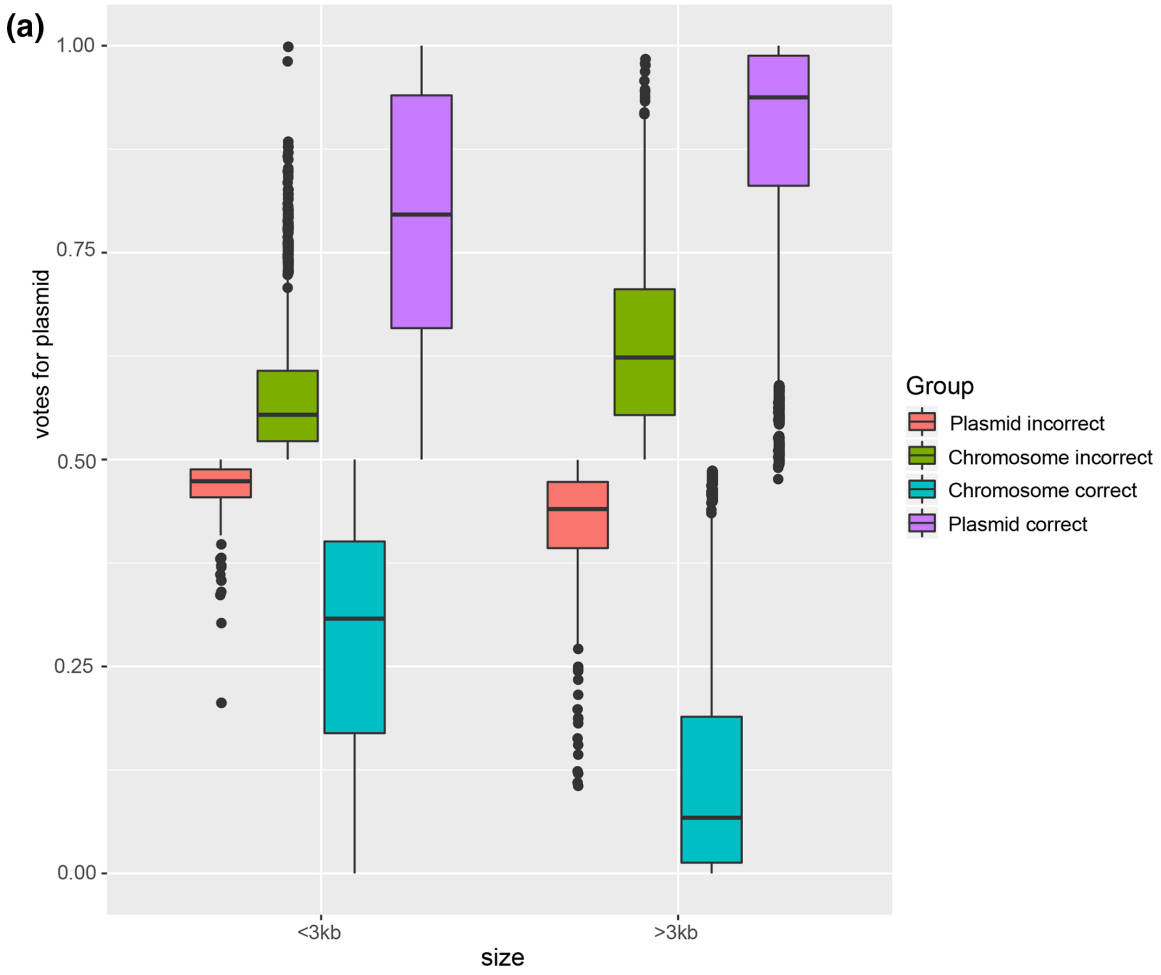
To address potential over-training, we present the OOB error and prediction failures of the complete model. Random Forest uses an internal validation where 66% of the contigs of the training sets are used for training and 33% are used for testing per tree in the Random Forest. The output of every tree is averaged and results in the OOB error, which is a minor over-estimation of the actual error [32]. The OOB classification results and the output of the complete model on the training data sets are presented in Fig. 2(a), Table S1. The results show that RFPlasmid can correctly identify the source of 87–100 % of the contigs, which is 95–100 % of the total bp count. Often, the taxon-specific model outperforms the taxon agnostic model (generic) (Fig. 2).

Random Forest outputs votes for the plasmid class (votes plasmid) and for the chromosomal class (votes chromosomal), ranging from 0 (=negative) to 1 (=positive). We observe that contigs that with scores between 0.4 and 0.6 are the main source of incorrectly predicted contigs (Fig. 3a). The incorrectly predicted contigs are mostly small contigs as shown in Fig. 3(b). Contigs smaller than 3 kb are difficult to classify, their scores are generally lower, possibly because the *k*-mer content cannot be reliably determined and specific *k*-mer content is an important feature for RFPlasmid classification (Fig. S1), or the contigs do not contain coding sequences (CDSs), whereas chromosomal and plasmid marker genes are also an import classification feature (Fig. S1). Furthermore, the small contigs consist of genes that usually have multiple copies on both genome and plasmid, such as transposases or phage genes [38]. To investigate the latter hypothesis, we determined the presence of phage genes and transposases on the incorrectly and correctly predicted contigs, and determined the phage and transposase content per contig. This analysis was performed on contigs containing at least one CDS. The highest rates of phage genes were found in the chromosome incorrectly classified contigs where 10% (1179 of 11 565) of the chromosome incorrect contigs consisted of >50% phage genes, compared to 6% (1316 of 22 063) of plasmid correct, 5% (5386 of 101 005) of chromosome correct and 3.5 % (13 of 372) of plasmid incorrect contigs (Fig. 4a). The highest rates of transposases were also found in chromosome incorrect contigs, where 22% (2514 of 11 565) of the chromosome incorrect contigs consisted of >50% transposases, compared to 14% (3125 of 22 063) of plasmid correct, 3% (3268 of 101 005) of chromosome correct and 7% (25 of 372) of plasmid incorrect contigs (Fig. 4b); and in the chromosome incorrect contigs, 59% (2487 of 4200) of the transposase-carrying contigs were small contigs (<3 kb).

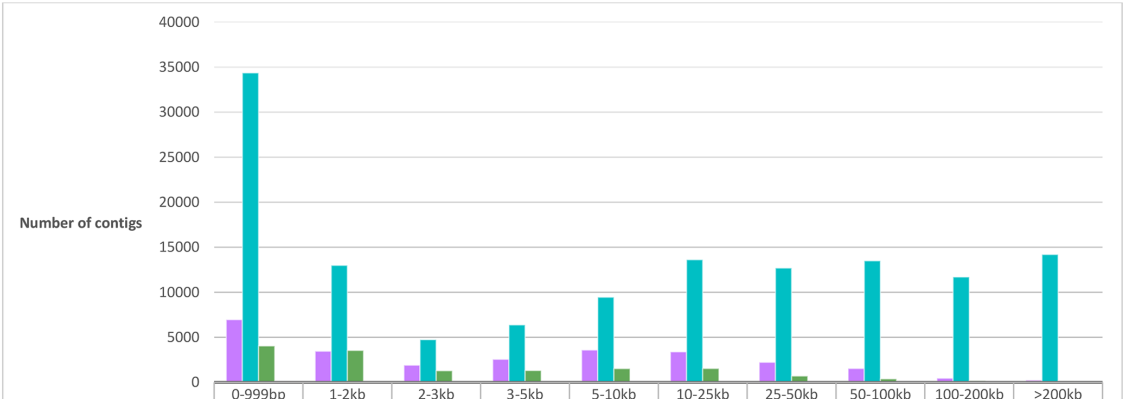
As the primary reason for our tool is to determine whether we can reliably predict whether AMR genes are carried on plasmids or chromosomes, we analysed the assembled contigs for the presence of resistance genes using the Resfinder database. Resistance genes were found on 5019 of the 175 027 contigs (135 004 contigs with >1 CDS) (Fig. 4c), of which 13% (2773 out of 21 306) of plasmids contigs carry an AMR gene and 1.77% (1977 out of 112 006) of the chromosomal contigs carry AMR genes. Only 3 out of 5019 AMR-harbouring contigs were plasmid incorrect contigs, and 4.3% (215 out of 5019) were chromosome incorrect contigs. Of these 213 chromosome incorrect AMR gene harbouring contigs, 38% (*n*=82) were located on small contigs (<3 kb); therefore, we conclude that we can reliably identify the DNA source that carries these genes, for example, for risk assessment.

Investigating the importance of each feature in the different training models shows that single-copy chromosomal markers genes and plasmid marker genes appear to function taxa wide as they are important in every model, while *k*-mer content is specific per taxon (Fig. S1). The specific *k*-mer content of each taxon is likely due to the correlation of the G+C content of plasmids with their host organism [39], where the plasmids have a lower G+C content compared to their hosts (Fig. S2) [40].

Batch processing of RFPlasmid is recommended, since the execution time of RFPlasmid is 1671 min for the bacteria model, consisting of 6895 files with a total of 1.19×10^{10} bp, which is a mean of 14 s per file by using 16 cores. The prediction of one single *Campylobacter* genome (ca. 2 Mbp) by using one core takes almost 8 min.



(b)



	0-999bp	1-2kb	2-3kb	3-5kb	5-10kb	10-25kb	25-50kb	50-100kb	100-200kb	>200kb
plasmid correct	6955	3459	1908	2551	3586	3382	2219	1542	458	247
chromosome correct	34355	12990	4744	6372	9453	13621	12686	13492	11691	14194
chromosome incorrect	4039	3539	1286	1309	1526	1542	702	385	187	119
plasmid incorrect	125	67	16	24	63	69	42	19	34	29

Fig. 3. RFPlasmid prediction results stratified for contig sizes. (a) Box plot displaying the plasmid prediction scores (votes_plasmid) of small (<3 kb) and large (>3 kb) contigs, grouped per correctly and incorrectly classified plasmid and chromosome contigs. (b) Graph of RFPlasmid prediction results, grouped per correctly and incorrectly classified plasmid and chromosome contigs, subdivided according to contig size.

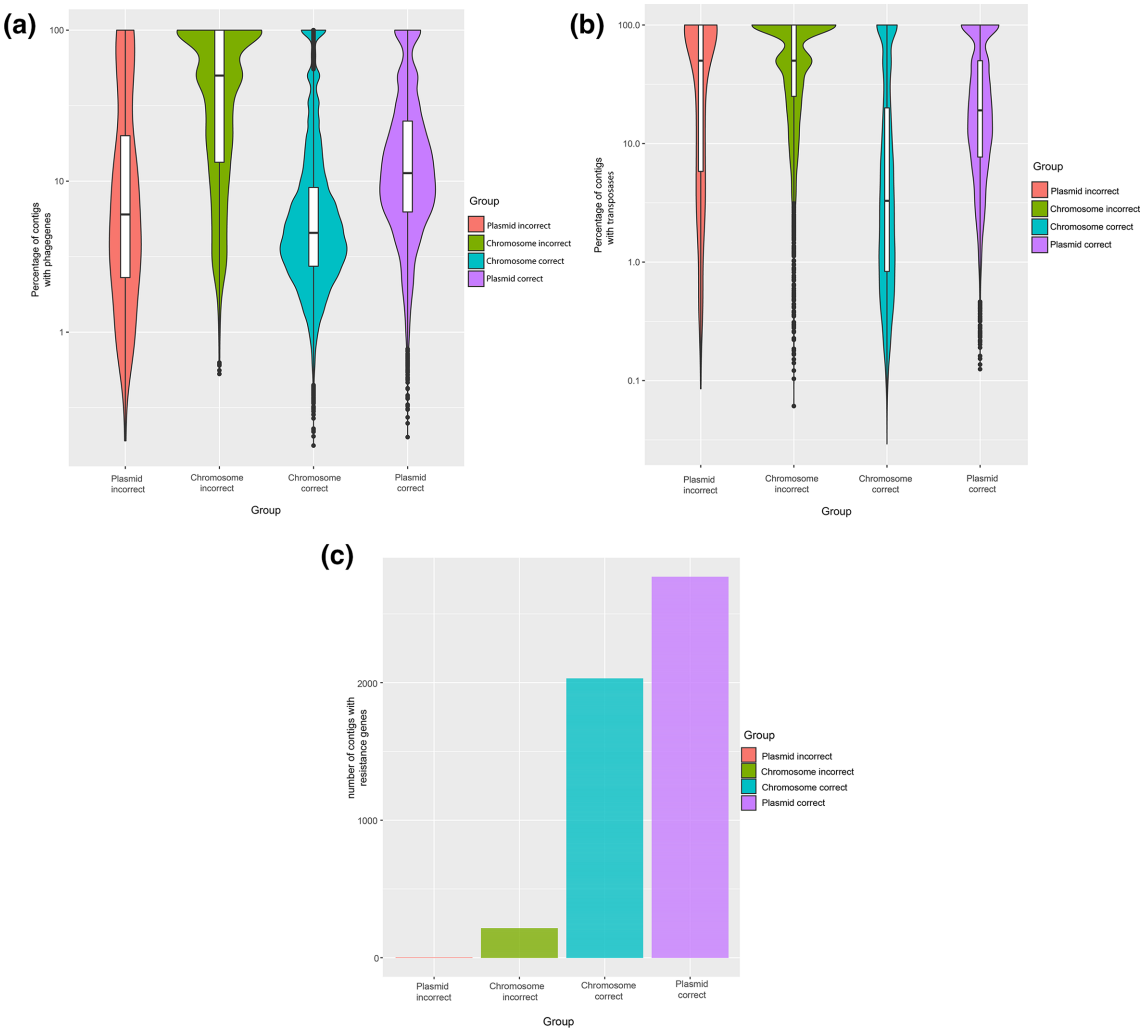


Fig. 4. Presence of phage genes, transposases and resistance genes in training data contigs. (a) A violin graph with box plot with the percentage of phage genes (log_{10} scale) in training data contigs, (b) a violin graph with box plot with the percentage of transposases (log_{10} scale) in training data contigs, and (c) bar plot with counts of contigs with ≥1 resistance gene, all grouped per correctly and incorrectly classified plasmid and chromosome contigs.

Benchmarking RFPlasmid and comparison with existing tools

We compared the performance of RFPlasmid with other plasmid-prediction tools. Plasmid-predictions tools that assemble plasmid contigs from read files like PlasmidSPAdes [15], Recycler [16] and PLACNET [18] are not developed to be used with assembled data and, therefore, are excluded from this comparison. The plasmid-prediction tools that can predict plasmid contigs from assembled data were tested and compared with RFPlasmid. The comparison was performed by using the models and training data sets described in this study: cBar [14] with the metagenome training data, PLAScope [19] with the *E. coli* subset of the *Enterobacteriaceae* training data, MLPlasmids [20] with the *Enterococcus faecium* and *E. coli* subsets of the *Enterococcus* training data and *Enterobacteriaceae* training data, respectively, Platon [21] with all taxa-specific models, the metagenome training data and the external *E. coli* set. Percentages of correctly predicted bp are calculated and compared with the RFPlasmid prediction results (Fig. 5, Table S2). We show that RFPlasmid outperforms the tested tools by having a lower number of incorrectly classified plasmid and chromosome contigs compared to cBar and MLPlasmids for *Enterococcus*, and by predicting a lower number of plasmid incorrect classified contigs compared to PLAScope. RFPlasmid outperforms Platon by having a higher number of correctly classified plasmid contigs [e.g. for taxon-specific models 26307 (RFPlasmid) vs 15257 (Platon) contigs], whereas Platon shows a slightly better prediction of chromosomal contigs compared to RFPlasmid [e.g. for taxon-specific models 133598 (RFPlasmid) vs 147 036 (Platon) contigs] (Table S2). RFPlasmid has a mean chromosome incorrect classified contig rate of 1.24% bp and a mean plasmid incorrect classified contig rate of 0.29% bp.

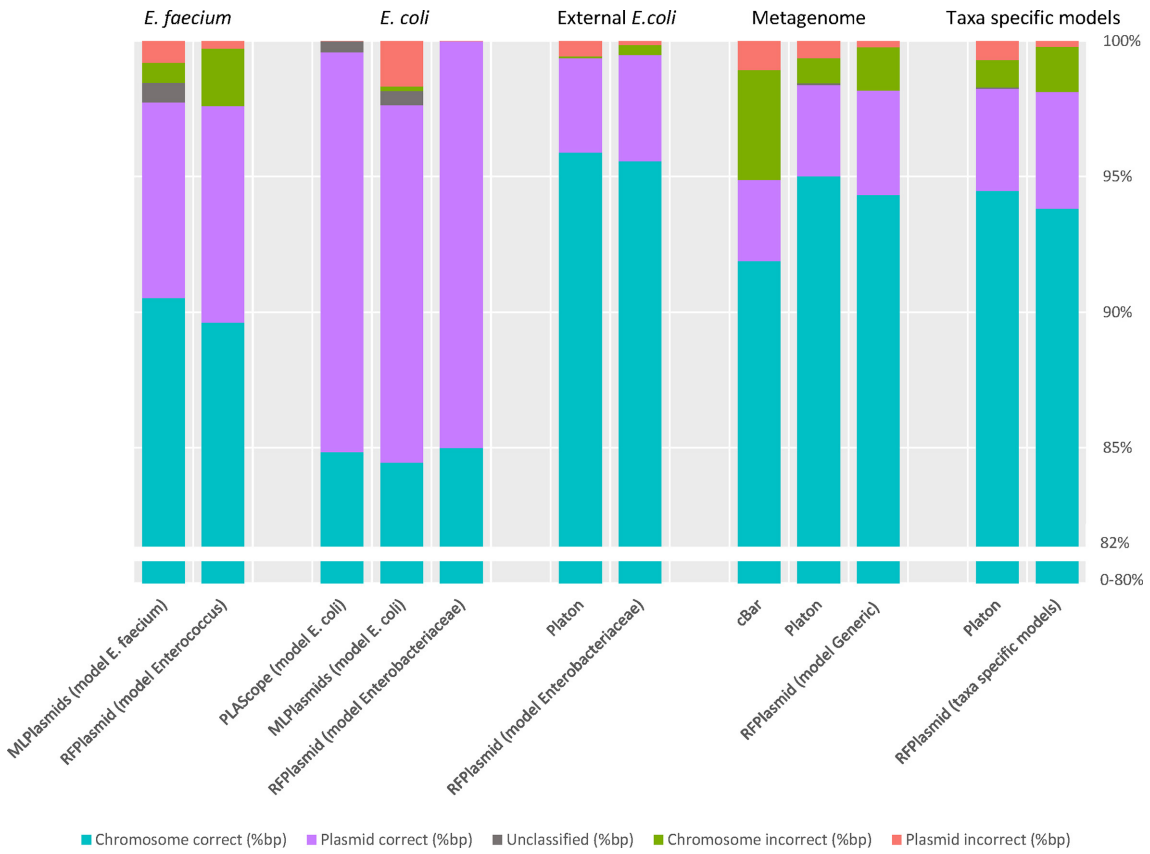


Fig. 5. Comparison of RFPlasmid performance with existing tools. Shown are the prediction performance of the compared tools for each specific model and associated training data set, represented in percentages (calculated as bp predicted divided by the total bp for each plasmid correct, chromosome correct, chromosome incorrect and plasmid incorrect contig). The *y*-axis is modified and starts from 82%, since percentages 0–80 % are all chromosome correct performances.

To investigate the performance of RFPlasmid on non-simulated data, we also used 22 *E. coli* genomes (external *E. coli* set), previously used by Schwenger *et al.* [21]. The error rate of RFPlasmid with non-simulated data is very low; only 0.52% of bp (85 contigs out of 2832 contigs) were incorrectly predicted with most of them (62 contigs out of 85 contigs) being small (<3 kb) (Fig. 5). Manual investigation of the larger incorrectly predicted contigs shows that 16 contigs contain phage-encoding genes and 3 contigs a plasmid replication gene, of which one encodes IncQ1, which is presumably integrated into the genome of isolate H69.

DISCUSSION

Identification of plasmid and chromosomal sequences is essential in surveillance of bacteria with plasmid-associated AMR and provides fundamental knowledge for molecular epidemiology and risk assessment of these bacteria. We showed that RFPlasmid is able to predict chromosomal and plasmid contigs with error rates ranging from 0.002 to 4.66 % (Fig. 2a) and that the use of taxon-specific models can be superior to a general plasmid prediction model. Single-copy chromosomal marker genes, plasmid genes, *k*-mer content and length of contig all appear to be informative; however, *k*-mer content is highly specific for taxa. Prediction of small contigs remains unreliable, since these contigs consists primarily of repeated sequences present in both plasmid and chromosome, e.g. transposases, or because *k*-mer content or marker genes cannot be easily identified. Contig length and inclusion of marker genes can also be influenced by the presence of repetitive sequences in the contig, which will increase the change of mis-assemblies. Repetitive sequences will also have reduced unique *k*-mer content, which make them harder to characterize. To solve the problem with small contigs that are part of larger plasmids, long-read sequencing can be a solution to obtain the complete sequence of the plasmids [33].

Comparison with existing methods shows that RFPlasmid generally performs equally or better to currently available methods. RFPlasmid is, to our knowledge, the first described tool that is optimized for 17 bacterial taxa and also includes a generic model when the taxon is not in the database (e.g. also suitable for metagenomics assembly data). If a good reference set with well identified chromosomal and plasmid contigs of another bacterial taxon is available, an easy training option is implemented in RFPlasmid, to train a new model

for this bacterial taxon. Our web-interface makes RFPlasmid accessible to users who are unfamiliar with the command line interface, which will improve uptake of the use of our tool.

Improvements are still possible for RFPlasmid. Careful examination of the incorrectly classified contigs shows that these frequently contain many phage genes or transposases. Phages are often found on chromosomes, rarely on plasmids; therefore, including a phage detection algorithm could certainly improve predictions, although that is out of scope for this study, as phage prediction has its own difficulties and complexities [41]. Furthermore, phage-like plasmids have been detected [42, 43] that would need to be investigated to see whether it is possible to distinguish these from real phages. Smaller contigs that consisting solely of transposases (1–3 kb usually) are generally present on both chromosome and plasmid, and these could be detected and marked as such. Integrated plasmids, such as the IncQ1 plasmid in the external dataset in *E. coli* isolate H69, show that some predictions will remain difficult. Other improvements could be the detection of rRNA operons, as these are usually chromosomally encoded, or circularization detection for revealing smaller plasmids [21]. An evaluation of the combination of the above-mentioned features with taxon-specific models would be interesting for future research.

Availability and requirements

Project name: RFPlasmid.

Project home page: https://github.com/aldertzomer/RFPlasmid.

Operating system(s): Linux (shell).

Programming language: Python, R, Bash.

Other requirements: CheckM, DIAMOND.

Optional: JELLYFISH.

License: e.g. AGPL.

Any restrictions to use by non-academics: none.

Funding information
This work received no specific grant from any funding agency.

Acknowledgements
The authors would like to thank Alex Bossers (Institute for Risk Assessment Sciences, Utrecht University, The Netherlands), Samuel Bloomfield (Quadram Institute, UK), Alison Mather (Quadram Institute, UK), Sophia Kathariou (North Carolina State University, USA) and Melissa Jansen van Rensburg (University of Oxford, UK) for contribution to the databases and testing the software, and Robert A. Petit for volunteering to build the Bioconda package.

Conflicts of interest
The authors declare that there are no conflicts of interest

References

1. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. *Microbiol Mol Biol Rev* 2010;74:434–452.
2. Dib JR, Wagenknecht M, Farías ME, Meinhardt F. Strategies and approaches in plasmidome studies – uncovering plasmid diversity disregarding of linear elements. *Front Microbiol* 2015;6:00463.
3. Li Y, Canchaya C, Fang F, Raftis E, Ryan KA, *et al*. Distribution of megaplasmids in *Lactobacillus salivarius* and other lactobacilli. *J Bacteriol* 2007;189:6128–6139.
4. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, *et al*. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother* 2018;73:1121–1137.
5. Carattoli A. Resistance plasmid families in Enterobacteriaceae. *Antimicrob Agents Chemother* 2009;53:2227–2238.
6. Johnson TJ, Nolan LK. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol Mol Biol Rev* 2009;73:750–774.
7. Sengupta M, Austin S. Prevalence and significance of plasmid maintenance functions in the virulence plasmids of pathogenic bacteria. *Infect Immun* 2011;79:2502–2509.
8. Goessweiner-Mohr N, Arends K, Keller W, Grohmann E. Conjugation in Gram-positive bacteria. *Microbiol Spectr* 2014;2:2.4.19.
9. Oniciuc EA, Likotrafiti E, Alvarez-Molina A, Prieto M, Santos JA, *et al*. The present and future of whole genome sequencing (WGS) and whole metagenome sequencing (WMS) for surveillance of antimicrobial resistant microorganisms and antimicrobial resistance genes across the food chain. *Genes* 2018;9:268.
10. Park SE, Pham DT, Boinett C, Wong VK, Pak GD, *et al*. The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nat Commun* 2018;9:5094.
11. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. *PLoS Genet* 2018;14:1–13.
12. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
13. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, *et al*. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* 2017;45:D535–D542.
14. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 2010;26:2051–2052.
15. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, *et al*. PlasmidSPAdes: Assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016;32:3380–3387.
16. Rozov R, Kav AB, Bogumil D, Shterzer N, Halperin E, *et al*. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics* 2017;33:475–482.
17. Carattoli A, Zankari E, Garciá-Fernández A, Larsen M, Lund O, *et al*. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother (Bethesda)* 2014;58:3895–3903.
18. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, *et al*. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet* 2014;10:12.
19. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, *et al*. PlaScope: A targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* 2018;4:1–8.
20. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, *et al*. Mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* 2018;4:11.
21. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, *et al*. Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein-sequence-based replicon distribution scores. *BioRxiv* 2020.
22. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:10.

23. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.
24. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
25. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, *et al*. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
26. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4.
27. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26:2460–2461.
28. Buchfink B, Xie C, Huson D. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
29. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27:764–770.
30. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002;2:18–22.
31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
32. Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. *PLoS One* 2018;13:e0201904.
33. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
34. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;31:3350–3352.
35. Hamada M, Ono Y, Asai K, Frith MC, Hancock J. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics* 2017;33:926–928.
36. Reis-Cunha JL, Bartholomeu DC, Manson AL, Earl AM, Cerqueira GC. ProphET, prophage estimation tool: A standalone prophage sequence prediction tool with self-updating reference database. *PLoS ONE* 2019;14:1–9.
37. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* 2017;33:3340–3347.
38. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev* 2018;31:1–61.
39. Almpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb Genom* 2018;4:0–7.
40. Rocha C, Danchin A. Base composition bias might result from competition for. *Trends Genet* 2002;18:291–294.
41. Arndt D, Marcu A, Liang Y, Wishart DS. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Brief Bioinformatics* 2018;20:1560–1567.
42. Galetti R, Andrade LN, Varani AM, Darini ALC. A phage-like plasmid carrying bla KPC-2 gene in carbapenem-resistant *Pseudomonas aeruginosa*. *Front Microbiol* 2019;10:2–6.
43. Octavia S, Sara J, Lan R. Characterization of a large novel phage-like plasmid in Salmonella enterica serovar Typhimurium. *FEMS Microbiol Lett* 2015;362:1–9.