PlasmidEC: An ensemble of classifiers that improves plasmidome recall from short-read sequencing data in *Escherichia coli*

Abstract

Over the past decades, pathogenic lineages of *Escherichia coli* have rapidly acquired antibiotic resistance. Currently, multidrug resistant *E. coli* is the most frequent cause of lethal infections among resistant bacteria in a hospital setting.[1] Antibiotic resistance genes (ARGs) are commonly spread via plasmids. From a clinical and epidemiological standpoint, it is very relevant to analyse the plasmid content in *E. coli*. The rise of *Illumina* whole genome sequencing (WGS) has enabled fast large-scale analysis of the genomic content of bacteria. However, it is usually not possible to reconstruct plasmids by genome assembly of short-read sequencing data. Therefore, several bioinformatic tools have been developed to uncover the total plasmid content in a sample, also referred to as the plasmidome, by classifying genomic sequences as either chromosome- or plasmid-derived. We benchmarked four of these binary classifiers (mlplasmids, PlaScope, Platon and RFPlasmids). They are at the basis of plasmidEC, an ensemble classifier that combines the output of three plasmid classifiers using a majority voting system. The combination of Platon/PlaScope/RFPlasmid presented the best plasmidome predictions (F1-score = 0.904). Compared to individual classifiers, plasmidEC achieved increased recall (0.885), especially for contigs derived from ARG-plasmids (recall = 0.941). Moreover, a plasmidome study of *E. coli* ST131 using plasmidEC was used to identify differences between this lineage and other *E. coli*. Finally, we show that plasmidEC removes chromosomal contamination in plasmid reconstructions obtained by MOB-suite.

Plain language summary

Bacteria often carry plasmids, which are small genetic elements that can be exchanged via horizontal gene transfer. In this way they can spread quickly between bacteria of the same or different species. Plasmids usually contain genes that allow their host to adapt to a specific environment. In pathogens, these include genes that provide resistance to antibiotics. Infections by antibiotic resistant bacteria are more difficult to treat, especially when they are resistant to multiple antibiotics. Multidrug resistance is becoming increasingly common in pathogenic lineages of *Eschericia coli*. For studying resistance in this species, accurate prediction of plasmid sequences is very important. Here, we use an *E. coli* dataset to compare four softwares which predict the total of sequences that originate from plasmids in a sample, i.e. the plasmidome. We developed plasmidEC, a tool that combines the predictions of three input softwares and outputs the prediction given by the majority of the tools. PlasmidEC improves the recall of plasmidome predictions, especially for plasmids that carry antibiotic resistance genes. We show two applications of this tool; plasmidome analysis of *E. coli* ST131, and removal of chromosomal contamination in reconstructions of individual plasmids.

Introduction

Antibiotic resistance is a major global health threat. The rapid emergence and spread of multidrug resistance is especially alarming, leading to infections that are complicated or even impossible to treat.[2] *Escherichia coli*, a commensal gut bacterium, has recently established successful clinical lineages due to the accumulation of antibiotic resistance genes (ARGs).[3,4] In 2019, 57.1% of *E. coli* clinical isolates from the European Union displayed resistance to at least one major antibiotic class, and 39.1% of these strains were resistant to multiple classes.[5] Third-generation cephalosporin-resistant *E. coli* has been estimated to account for 47.4% of total infections and 32.1% of attributable deaths caused by resistant bacteria in 2015. The number of attributable deaths by this resistance group, which commonly produces extended-spectrum β-lactamases (ESBLs), was estimated to have increased at least four-fold since 2007. [1]

Plasmids play a critical role in the dissemination of antibiotic resistance.[6] These independent genetic elements usually contain genes that contribute to the adaptation of bacteria to specific environments, and they can spread within and between species by diverse horizontal transfer mechanisms.[7,8] ARGs to all major antibiotic classes have been found on plasmids, including those encoding ESBLs, carbapenemases and quinolone resistance. [9] Therefore, it is very relevant to study the plasmid content of *E. coli* and its impact on the spread of resistance.

Traditionally, plasmid sequences were determined by purifying plasmid DNA in the laboratory, followed by shotgun sequencing. However, this method is labour intensive and not applicable for recovering large plasmids.[10,11] Recently, *Illumina* short-read sequencing has become the most popular technology to determine the genomic content of bacteria. This method is accessible, fast and allows the processing of many samples at once. However, due to the frequent occurrence of repeated elements, it is usually not possible to assemble complete plasmids using short-read sequencing data alone. Instead, genome assemblies result in hundreds of contigs of unclear origin (plasmid or chromosome).[12] A hybrid approach, which uses both long-read and short-read sequencing information, is able to resolve this issue but is also costly. [13] Several bioinformatic tools have been developed to predict the plasmid content of bacteria from short-read sequencing data. According to their function, these softwares can be broadly categorised into two classes: binary classifiers, which classify assembled contigs as either chromosome- or plasmid-derived, and plasmid reconstruction softwares, which aim at reconstructing individual plasmids.

Binary classifiers can distinguish between chromosomal and plasmid contigs by exploiting differences in k-mer content, aligning contigs to reference databases or detecting plasmid-specific genes. These tools output the entire plasmid content of a strain, also referred to as the plasmidome. Plasmidome analysis has proven very useful to uncover the genomic location of ARGs [14–17], and also to determine the role of the plasmidome in niche adaptation[18]. There exist several binary classifiers, using different computational strategies, that can be used to identify the plasmidome of *E. coli*. However, as of to date the performance of these tools has not been evaluated in an independent manner. Besides, there is a need to assess their suitability for uncovering contigs of plasmids that carry ARGs, which are of specific clinical relevance. Furthermore, a major problem of these classifiers is that they suffer from low recall and may be biased towards recovering only a certain type of plasmid.[19–22]

In this work, we compare the performance of four binary classifiers using a comprehensive and diverse dataset of *E. coli* genomes. We present plasmidEC, an ensemble classifier that implements a majority voting system based on the combined output of three individual classifiers. We show that plasmidEC provides plasmidome predictions with increased recall, especially for contigs derived from ARG-plasmids. Moreover, plasmidEC was used to study the plasmidome and resistome of a collection of *E. coli* strains obtained from patients treated in different ICUs across Europe. Our tool allowed us to identify plasmidome specific differences between ST131 and other STs. A common problem of plasmid reconstruction tools is that their predictions contain chromosomal sequences.[23] We show that plasmidEC refines individual plasmid reconstructions obtained with MOB-suite, by removing chromosomal contamination.

Materials & methods

All scripts used can be found at https://github.com/lisavader/ST131. R version 4.1.0 [24] was used for all R scripts.

1. Benchmarking study of binary classifiers and plasmidEC

Sample selection
A dataset of 240 *E. coli* complete genomes carrying 631 plasmids was selected from Paganini et al.[23] Samples were isolated from animals, humans and the environment, resulting in a diverse dataset with respect to phylogeny and plasmid content. All genomes were completed by hybrid assembly. Short-read sequences and completed genomes were downloaded from NCBI using ncbi-genome-download v0.2.10 (https://github.com/kblin/ncbi-genome-download/). Samples present in the training datasets or reference databases of mlplasmids, PlaScope, Platon and/or RFPlasmid were removed (n=26). One sample was removed due to difficulties during genome assembly. The final dataset consisted of 213 samples including 542 plasmids.

Selection of contigs for benchmarking
Short-read sequences of each sample were assembled using bactofidia v1.1. (https://gitlab.com/aschuerch/bactofidia). The resulting contigs (n= 18,963) were labeled as chromosome- or plasmid-derived by alignment to their respective genomes using QUAST v5.0.2.[25] Only contigs larger than 1000 bp with an alignment of at least 90% the contig length were included (n=15,020). Contigs aligning to multiple positions at the genome (ambiguously aligned contigs) were included as long as they exclusively aligned to either the chromosome or to plasmids (n=1,236). The same applies for contigs that partly align to one position, and partly to another (misassembled contigs) (n=1,862). In total, the benchmarking dataset included 14,746 contigs (Figure S1).

Assessment of binary classifier performance
Contigs were classified by mlplasmids v2.1.[20], PlaScope v.1.3.1[21]**,** Platon v.1.6[19] and RFPlasmid v.0.0.17[22]. All tools were run using default parameters. We assessed the performance of the four binary classifiers by comparing, for each contig, the binary prediction to their actual class, as previously determined by genome alignment. For PlaScope, an 'unclassified' prediction was handled as a negative prediction (chromosome). Predictions were categorised into True Positives (TP; prediction=plasmid, class=plasmid), True Negatives (TN; prediction=chromosome, class=chromosome), False Positives (FP, prediction=plasmid, class=chromosome) and False Negatives (FN, prediction=chromosome, class=plasmid). Each tool was evaluated with respect to recall [TP/(TP+FN)], precision [TP/(TP+FP)] and F1-score [2*(recall*precision)/(recall+precision)].

Assessment of ensemble classifier performance
Majority voting ensemble classifiers were tested using four different combinations of binary classifiers: mlplasmids/PlaScope/Platon, mlplasmids/PlaScope/RFPlasmid, mlplasmids/Platon/RFPlasmid and PlaScope/Platon/RFPlasmid. Results were calculated in R, based on the overlap between the predictions of binary classifiers. The ensemble classifiers were evaluated using the same metrics as described for the binary classifiers (recall, precision, F1-score).

Selection of ARG-plasmids
A subset of ARG-plasmids (n = 112) was selected from Paganini et al.[23] This dataset consists of plasmids that contain at least one ARG, as determined with ABRicate v1.0.1 (https://github.com/tseemann/abricate),using the ResFinder database.[26]

2. Plasmidome analysis of *E. coli* ST131

Genome assembly and taxonomic classification

Samples were sequenced by Illumina sequencing using the NexteraXT library preparation kit. Read lengths were 150 bp. Genome assembly was performed by bactofidia v1.1. Species and ST were determined by multi-locus sequence typing using mlst v.2.16.2 (https://github.com/tseemann/mlst). Phylogroups were determined by ClermonTyping v.20.03.[27] In total, 363 *E. coli* samples were detected by mlst. One of these samples was excluded because it was recognised as *E. marmotae* by ClermonTyping. The fimH allele of the samples was determined using blastn v.2.12.0+[28] against the FimH database from FimTyper v1.0 (https://bitbucket.org/genomicepidemiology/fimtyper/src).

Accessory genome and plasmidome analysis

Gene annotation was done with prokka v.1.14.5.[29] Core and accessory genome content were determined using Panaroo v.1.2.3 run in sensitive mode.[30] The plasmidome was determined using plasmidEC. Accessory genome and plasmidome Jaccard distances were calculated in R using proxy v.0.4.26.[31]

Phylogenetic reconstruction

A maximum likelihood phylogenetic tree was built using RAxML v.8.2.12[32] based on the core gene alignment produced by Panaroo. RAxML was run over 20 iterations using the CAT model of rate heterogeneity. The tree was rooted using the *E. marmotae* sample as outgroup.

Resistome analysis

We predicted ARGs present in our dataset by running ABRicate v.0.8 against the ResFinder database, using a coverage cut-off of 95%. The origin of contigs containing ARGs was predicted using plasmidEC. Cooccurence analysis of ARGs was carried out in R using the package cooccur v.1.3[33], visualisation was done by visNetwork v.2.1.0[34]

3. Removing chromosomal contamination in plasmid predictions

Reconstruction of plasmids

The dataset used is equal to the benchmarking dataset used to evaluate the binary classifiers. Plasmids were reconstructed by MOB-suite v.3.0.0.[35] In case of the 'reconstruction first' method, plasmidEC was run on the resulting predictions and all contigs predicted to be chromosomal were removed. In case of the 'selection first' method, contigs were first classified by plasmidEC and MOB-suite was then run using as input all plasmid-predicted contigs.

Evaluation of performance

The contigs of the predicted plasmid units (bins) were aligned to their completed reference genomes using QUAST v5.0.2. A length cut-off of 1000 bp and coverage cut-off of 90% were used for alignment. Ambiguous alignments (alignments of a contig to multiple reference replicons) were included, except when the bin was composed solely of ambiguously aligned contigs.

Bins were evaluated with respect to recall and precision. The recall per bin-plasmid alignment is defined as the fraction of the reference plasmid covered by the bin. The precision is defined as the fraction of the bin that aligns to the reference plasmid. The total recall for each reference plasmid was calculated by adding all alignments to that reference plasmid together.

Results

1. Benchmarking study of binary classifiers and plasmidEC

Selection of binary classifiers

We compared four binary classifiers that use distinct computational strategies for plasmid classification: mlplasmids, Platon, PlaScope and RFPlasmid. Mlplasmids uses support-vector machine models to distinguish contigs based on their pentamer frequencies. It has five species specific models available.[20] Platon classifies contigs based on the distribution of protein-coding genes, which is different for chromosomal and plasmid sequences. Furthermore, Platon uses specific sequences such as mobilisation genes, incompatibility sequences and oriT, and Blast hits to a plasmid database.[19] PlaScope performs k-mer searches against custom databases of plasmid and chromosome sequences, currently available for *E. coli* and *K. pneumoniae*. These searches are applied using the metagenomic classifier Centrifuge.[36] Contigs that don't have any hits, have hits to both plasmid and chromosome sequences, or don't have sufficient length or coverage are assigned to an 'unclassified' category.[21] RFPlasmid uses a random forest classifier based on k-mer composition, plasmid and chromosomal marker genes, replicons, overall gene content and contig length. It has models available for sixteen genera, the Enterobacteriaceae family and a general Bacteria model.[22]

Binary classifiers show major differences in performance

After genome assembly using short-reads, we obtained a total of 18,963 contigs, of which 77.8% (n=14,746) were included in the final benchmarking dataset . An overview of the included and excluded contigs and their alignment type can be found in Figure S1. Of the included contigs, 87.3% (n=12,872) were of chromosomal origin, while the remaining 12.7% (n=1,874) were plasmid derived. The class of all included contigs was predicted by the four selected binary classifiers. Predictions were later compared to the true class of the contigs, which was determined by aligning each contig to its corresponding complete genome. The performance of the tools was evaluated using the metrics precision, recall and F1-score.

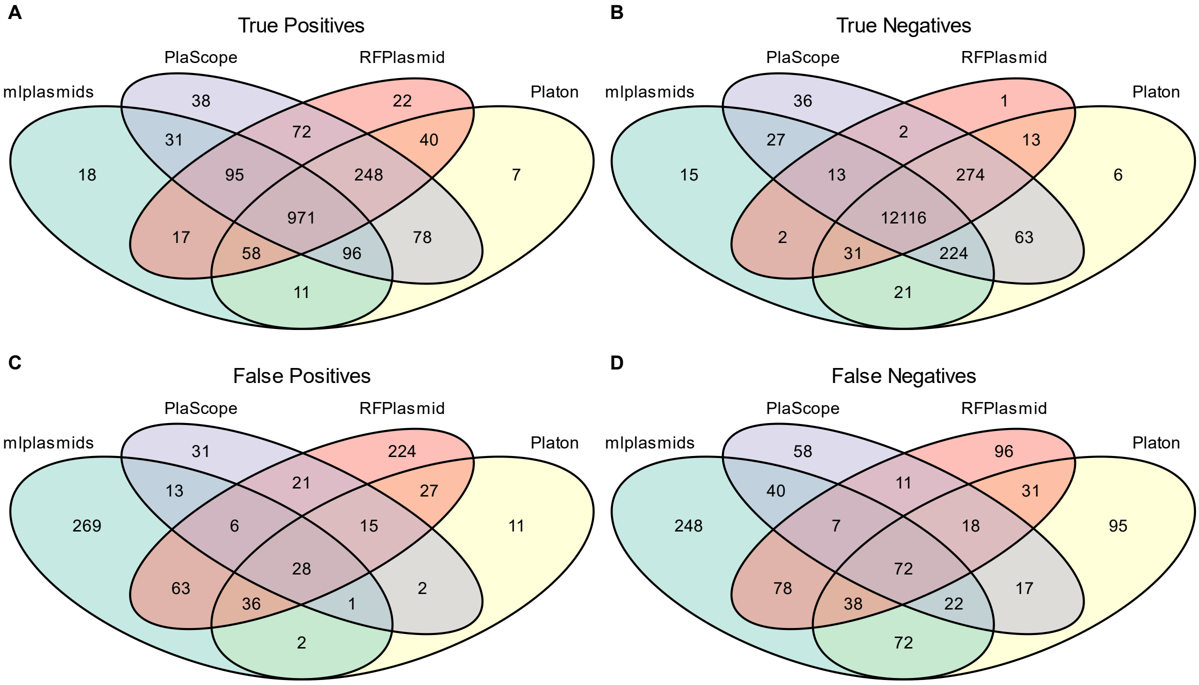


Figure 1 - Venn diagrams representing the overlap in absolute count of True Positives (a), True Negatives (b), False Positives (c) and False Negatives (d) between binary classifiers.

We examined the overlap in contig predictions between the tools. A high fraction of chromosomal contigs was correctly classified by all of the tools (n=12116, 94.1%) (Figure 1A), but this was only the case for approximately half of plasmid contigs (n=971, 51.8%) (Figure 1B). The majority of misclassifications, for both plasmid- and chromosome-derived contigs, were made by only one of the softwares (FP: 535, 71.4%, FN: 497, 55,0%) (Figures 1C and 1D). In contrast, a low percentage of correct predictions was unique to one software (TP: 85, 4.7%, TN: 58, 0.5%) (Figures 1A and 1B).

The four classifiers showed large differences in recall, precision and F1-score (Figure 2A). The best performance was reached by PlaScope, which presented the highest values for all metrics (recall = 0.869, precision = 0.933, F1-score = 0.900). Platon scored similarly in terms of precision (0.925), but achieved a lower recall (0.805). All softwares except RFPlasmid showed a higher precision than recall. Table S1 provides an overview of all results per software.

PlasmidEC improves the recall of contigs derived from ARG plasmids

We built plasmidEC, an ensemble classifier that combines the predictions of three individual classifiers and outputs the prediction given by the majority of the tools (Figure S2). The rationale behind this is to discard software-specific misclassifications, while keeping correct classifications, which are usually shared between softwares. Additionally, the combination of multiple classification methods could broaden the variety of plasmid sequences that can be detected. PlasmidEC is publicly available at https://github.com/lisavader/plasmidEC/.

We tested all possible combinations of input classifiers and their effect on recall, precision and F1-score (Figure 2B). The combination of Platon/PlaScope/RFPlasmid presented the best overall performance (recall = 0.885, precision = 0.924, F1-score = 0.904). This ensemble classifier achieved an F1-score similar to PlaScope, but recall and precision values were more balanced. Scores for all ensemble classifier combinations can be found in Table S2.

We also evaluated recall values of all individual and ensemble classifiers for a subset of plasmids (n=112) encoding antibiotic resistance genes (ARG-plasmids) (Figure 2C and 2D, Table S3). This dataset consisted of 860 plasmid contigs, from 91 *E. coli* genomes. Notably, all tools showed an increased recall when detecting contigs derived from these ARG-plasmids. However, for this dataset, the combination of Platon/PlaScope/RFPlasmid (recall =0.941) strongly outperformed the best individual classifier (Plascope, recall=0.884) . This major improvement indicates that plasmidEC is especially suited for identifying contigs from ARG-plasmids.

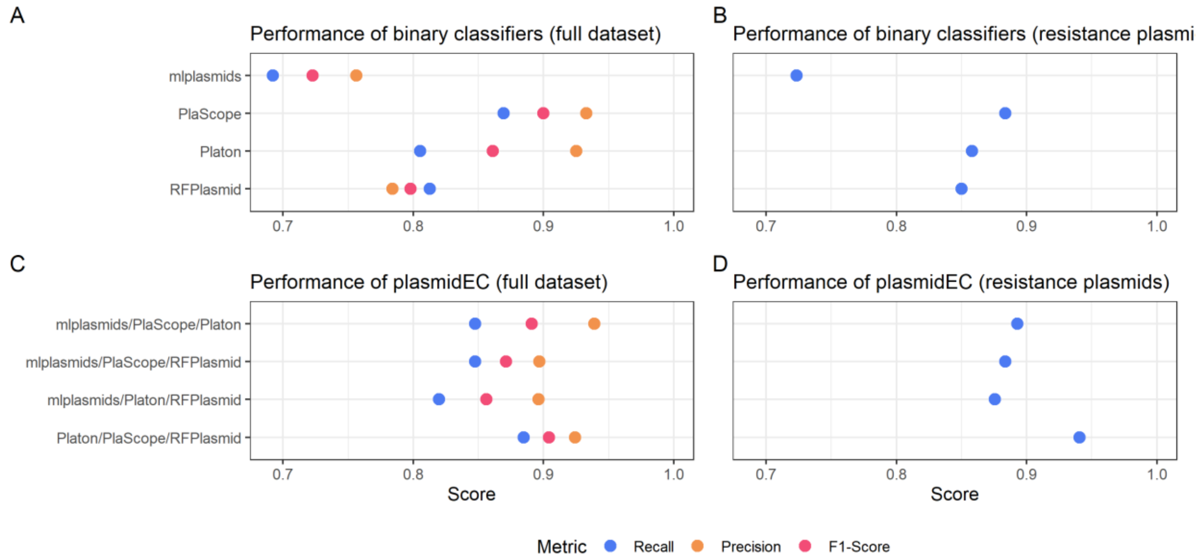


Figure 2 - Performance of binary classifiers and plasmidEC combinations, measured by recall (blue), precision (orange) and F1-score (pink) of contigs. (a) Binary classifiers evaluated using full dataset. (b) Binary classifiers evaluated using dataset of plasmids containing ARGs. (c) PlasmidEC combinations evaluated using full dataset. (d) PlasmidEC combinations evaluated using dataset of plasmids containing ARGs.

Computational performance

Finally, we measured the computational resources used by the individual and ensemble classifiers (Figure 3). The binary classifiers showed considerable differences in both CPU time and memory used. The average CPU time required per sample was lowest for PlaScope (0.2 mins) and highest for Platon (14.9 mins). Platon also used the largest amount of memory per sample (20.6 Mb). The least amount of memory was required by mlplasmids (2.7 Mb). Because plasmidEC includes the execution of three binary classifiers, time and memory requirements were high, especially when Platon was run. The combination of mlplasmids/PlaScope/RFPlasmid required the least amount of resources (CPU time = 4.5 mins, memory = 9.0 Mb) and PlaScope/Platon/RFPlasmid the most (CPU time = 21.5 mins, memory = 21.4 Mb).

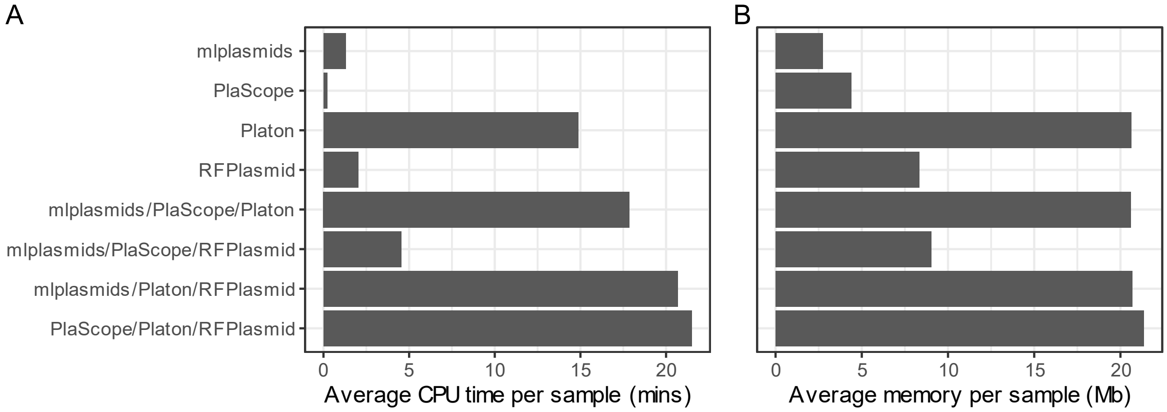


Figure 3 - Average computational resources used per sample: CPU time in minutes (a) and memory in Mb (b). Softwares were run on the full benchmarking dataset (n = 213) using 8 CPU’s.

2. Plasmidome analysis of *E. coli* ST131

E. coli ST131 is a pandemic pathogen and a frequent cause of urinary tract and bloodstream infections.[37] This clonal group was first detected in 2003, spreading very rapidly to become the most abundant sequence type (ST) among *E. coli* clinical isolates.[38] *E. coli* ST131 is commonly associated with fluoroquinolone resistance and with the production of ESBLs. We applied plasmidEC for exploring the plasmidome of a set of *E. coli* ST131 isolates, and compared it to other *E. coli* of diverse STs. We also studied and contrasted the entire accessory genome and resistome of this set of isolates.

Population structure of highly resistant *E. coli* in European ICUs

We analysed a dataset of 362 *E. coli* isolates that were collected from ICUs of 10 European hospitals between 2013 and 2017, as part of the RGNOSIS-ICU study, which examined the effect of decontamination strategies on the incidence of bloodstream infections in ICUs.[39] Most of the samples (330, 91.2%) were selected for ESBL production. Samples were isolated from the rectum (331, 91.4%), the respiratory tract (23, 6.4%) or the blood (8, 2.2%). Multi-locus sequence typing detected 94 different STs, showing the considerable phylogenetic diversity of our dataset. ST131 was the most abundant sequence type (n=98, 27.1%), followed by ST10 (n=20, 5.5%) (Figure S3). Phylogenetic diversity was confirmed through the construction of a maximum likelihood tree, based on core SNP distances between the samples (Figure 4a).

Within *E. coli* ST131, the majority of samples belonged to clade C (n=87, 88.8%), as determined by typing of the fimH allele, and only 10 samples corresponded to either clade A (n=7, 7,1%) or clade B (n=7, 7.1%).(Figure 4b). Among these isolates, we found 9 different blaCTX-M alleles, of which blaCTX-M-15 and blaCTX-M-27 were most frequent, occurring in 49 samples (50.0%) and 30 samples (30.6%) respectively. Interestingly, clade C isolates with the same blaCTX-M allele also presented a stronger phylogenetic association, as they clustered together in the maximum likelihood tree (Figure

4b). Despite this, we did not find the distinct subclades C1 and C2 associated with blaCTX-M-15 and blaCTX-M14/27 that have been reported previously.[40,41]

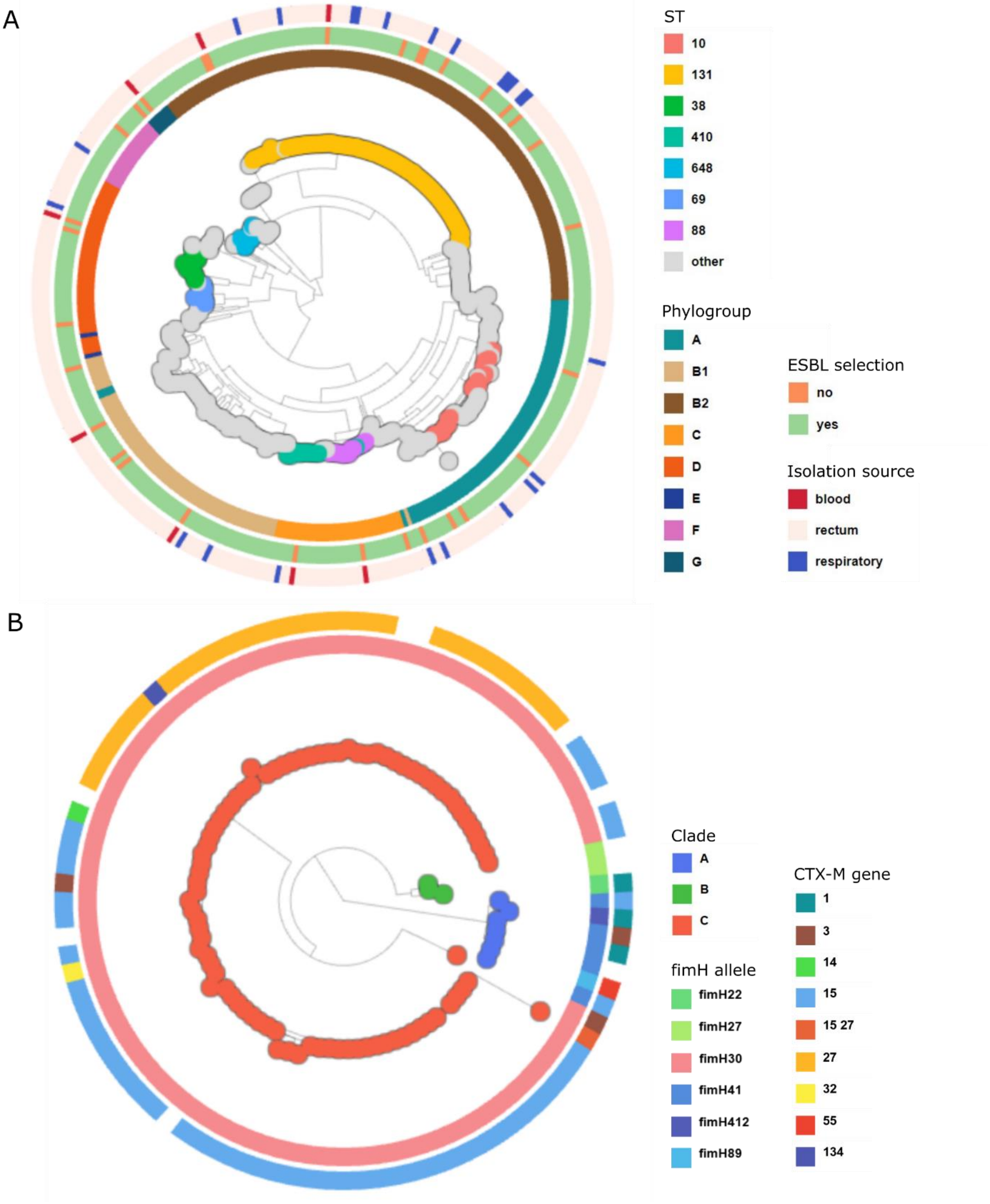


Figure 4 - (a) Maximum likelihood phylogeny of *E. coli* samples, tips are coloured by sequence type (ST, only STs present in at least 10 samples). Rings are coloured by phylogroup, selection for ESBL and isolation source. b) Maximum likelihood phylogeny of *E. coli* ST131 samples, tips are coloured by clade. Rings are coloured by fimH allele and blaCTX-M type.

E. coli ST131 presents a distinct accessory genome and plasmidome

We examined differences in the accessory genome and in the predicted plasmidome of the *E. coli* samples. Every sample in our dataset contained contigs predicted to originate from plasmids. On average, the accessory genome consisted of 2200 genes, and the plasmidome of 294 genes. We found a positive correlation between plasmidome and accessory genome size (Figure S4). Notably, we discovered significant differences in accessory genome and plasmidome size between *E. coli* ST131 and other STs. Compared to other lineages, *E. coli* ST131 carried a more expansive accessory genome (mean = 2310), but a more limited plasmidome (mean = 233) (Figure 5).

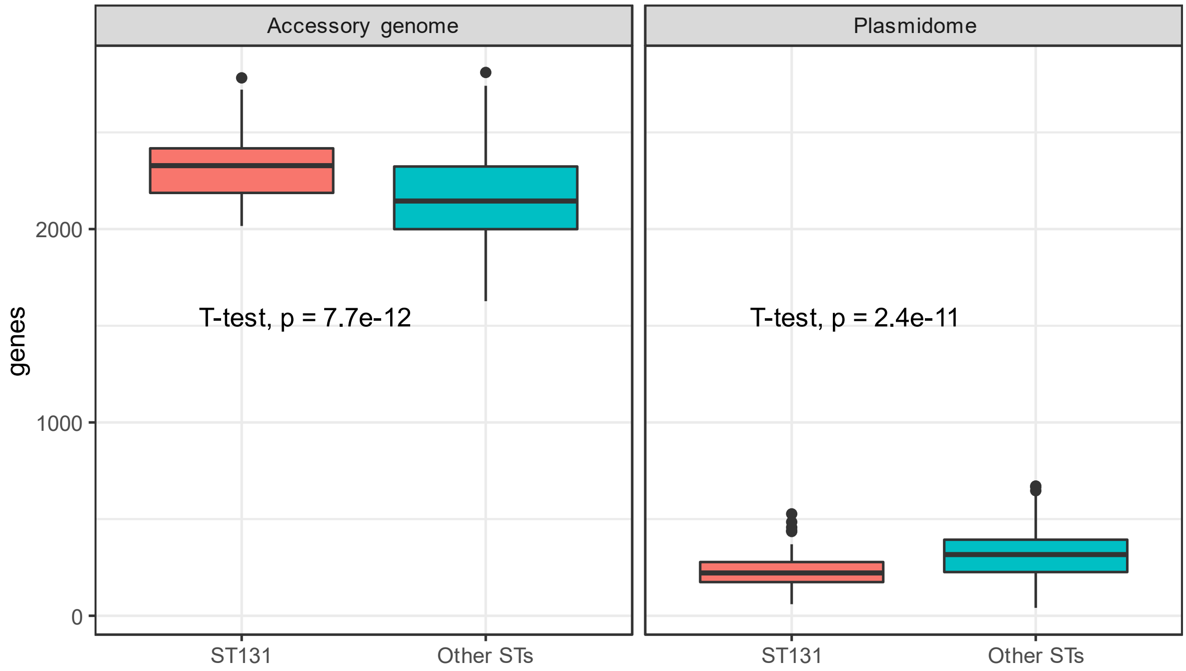


Figure 5 - Number of genes present in the accessory genome and plasmidome of *E. coli* ST131 (red) compared to other STs (blue). Significance determined by Welch t-test.

To evaluate the differences in accessory genome and plasmidome composition, we first obtained a gene presence-absence matrix and used this to calculate Jaccard distances between all samples. These distances were later given as an input for a hierarchical clustering algorithm (Figure 6). The accessory genome composition appears to have a strong association with the phylogeny of the isolates, especially at the phylogroup level. Figure 6 shows two main clusters, one composed of isolates that belong to phylogroup B2, and another containing a mix of phylogroups A, B1 and C, three phylogroups that share a common phylogenetic origin (Figure 4a). Within phylogroup D, the accessory genome seems to be related to ST but not to the overall phylogroup. All ST131 strains are part of a highly conserved cluster which is very distinct from other strains.

The plasmidome shows more variation than the accessory genome, and is not clearly associated with the phylogroup of the isolates. However, many of the *E. coli* ST131 strains still cluster together, indicating that there are shared plasmid sequences between them.

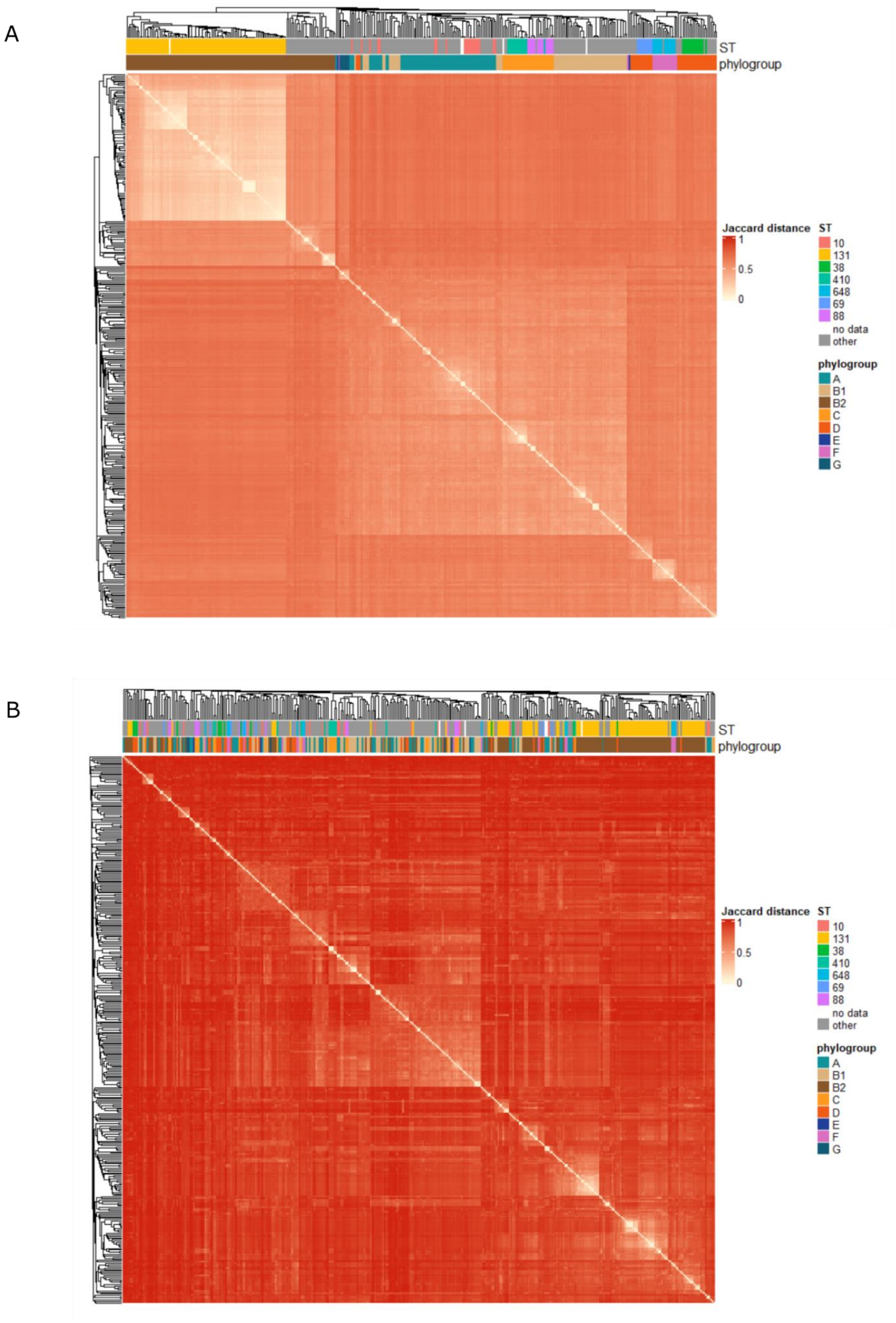


Figure 6 - Heatmaps showing the Jaccard distances between the accessory genome (a) and plasmidome (b) of samples, clustered using hierarchical clustering (linkage = complete). Samples are coloured by phylogroup and ST (only STs present in at least 10 samples).

E. coli ST131 carries a limited but specific resistome

E. coli plasmids encode most of its resistome, since the majority of ARGs (n=579, 82.4%) were found on plasmid predicted contigs (Figure S5). Notably, *E. coli* ST131 carries a significantly smaller number of ARGs (mean=7.2) than other STs (mean=8.4) (Figure S6). Additionally, ST131 samples showed limited variety in the resistome (Figure 7), encoding a total of 39 distinct ARGs and frequently displaying specific ARG combinations. In contrast, other *E. coli* displayed a more diverse resistome, encoding 110 distinct ARGs. This variety is also present for individual STs, for example ST10, which is represented by only 20 samples compared to 98 ST131 samples, encodes a total of 52 distinct ARGs.

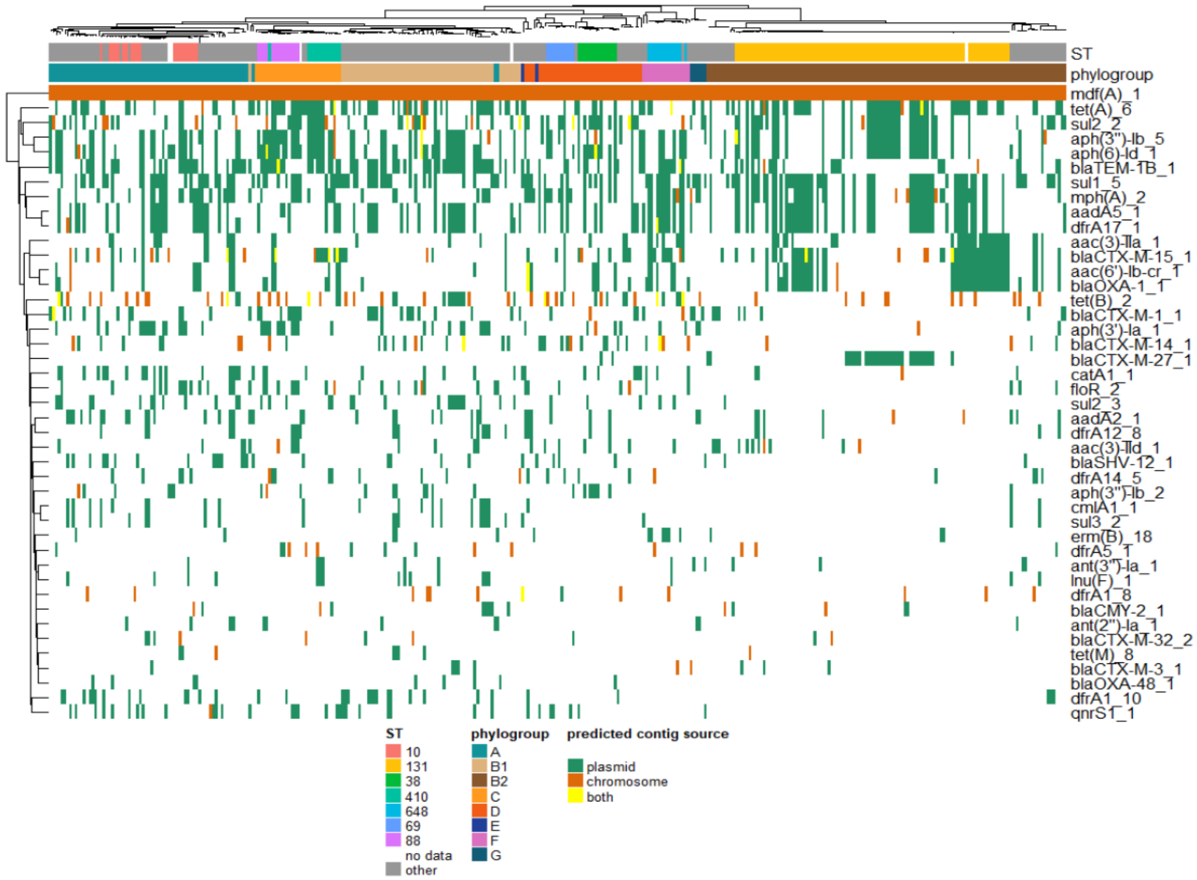


Figure 7 - Antibiotic resistance gene (ARG) profiles per sample. ARGs are coloured by their predicted location; plasmid (green), chromosome (orange) or both (yellow). Samples are clustered by phylogeny and coloured by phylogroup and ST (only STs present in at least 10 samples). Showing only ARGs present in at least 8 samples.

We analysed the cooccurrence of ARGs within samples to unravel clusters of resistance determinants. In *E. coli* ST131, we found four distinct clusters, which are annotated in Figure 8. Clusters 1 and 2 were previously described in plasmids associated with clade C1 and clade C2 respectively.[34] There is a negative correlation between the presence of cluster 1 and 2, and notably, both clusters contain different alleles of a blaCTX-M gene. Interestingly, tet(A) appears to be the main resistance determinant against tetracycline in ST131. This ARG is associated with both cluster 3 and 4, two clusters composed of genes that are not significantly associated with each other, and that show redundant functions. Within other *E. coli*, significant relationships between ARGs are more abundant and more complex. Nevertheless, we also observe mutual exclusivity between genes that encode a common resistance type. This mutual exclusivity is present for the different blaCTX-M genes, as well as for sul and tet genes, which encode sulfonamide and tetracycline resistance respectively.

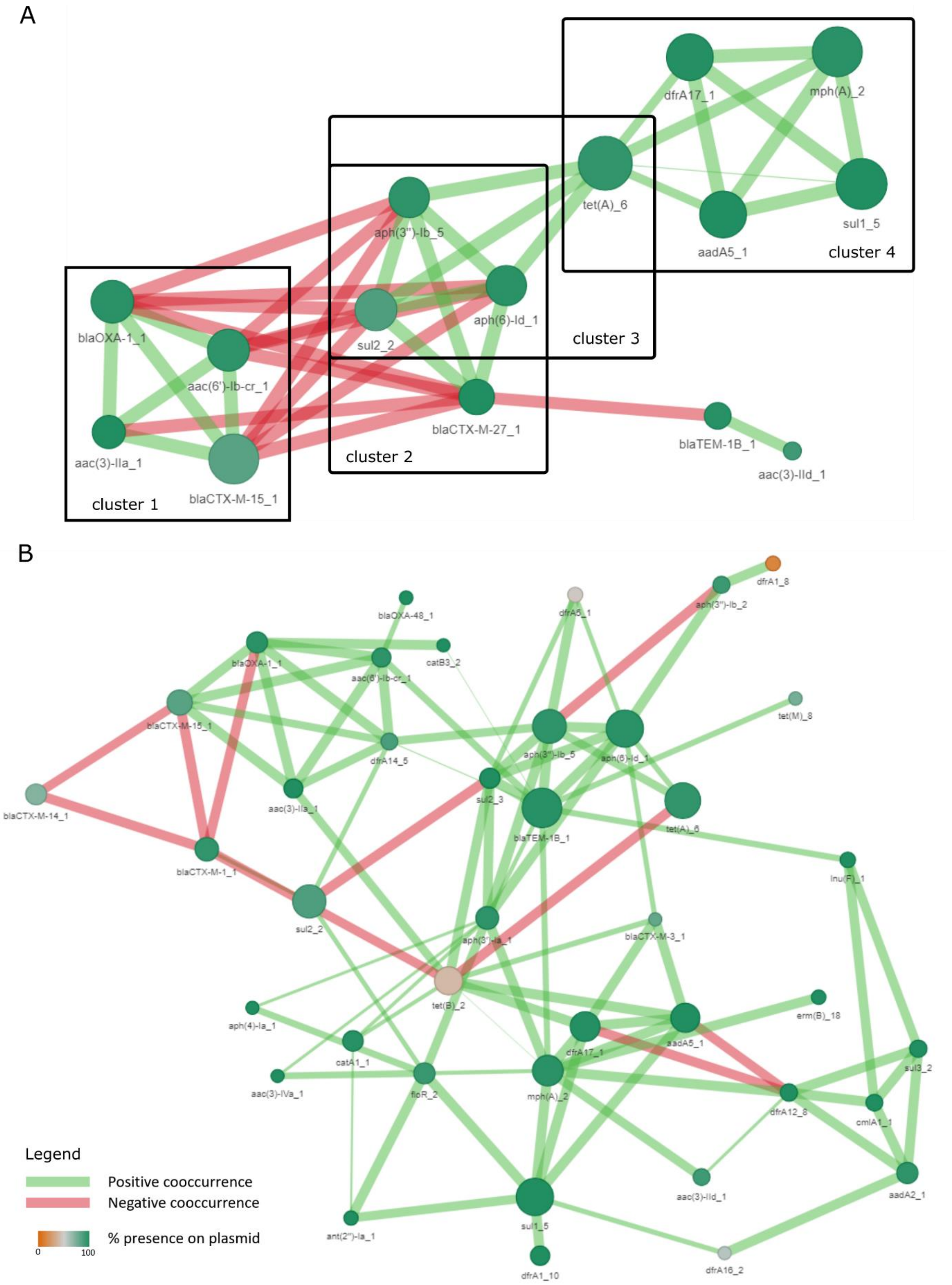


Figure 8 - Cooccurrence networks of ARGs found in *E. coli* ST131 (a) and in other STs (b). Showing only significant positive (green) and negative (red) interactions. Thicker edges correlate with lower p-values. Node size corresponds to abundance and node colour corresponds to the average predicted presence on plasmid vs. chromosome. ARG clusters are annotated by hand.

3. Removing chromosomal contamination in plasmid predictions

In a recent work, MOB-suite presented the best performance for reconstructing individual *E. coli* plasmids from short-read data (50.2% correct reconstructions). [23] A major flaw of this software, however, was the inclusion of chromosomal chromosomal contigs in a significant fraction of the plasmid predictions (40%). We evaluated the performance of plasmidEC as a tool for removing chromosomal contamination from MOB-suite predictions.

Tools were applied to the same dataset that was used for the benchmarking of binary classifiers. To test the influence of the assembly step, short-reads were assembled using two different pipelines: Unicycler[13] and bactofidia[42]. The resulting contigs were used as input for MOB-suite and plasmidEC. We tested two different methods for integrating the tools. According to the 'reconstruction first' method, plasmids are reconstructed by MOB-suite, after which chromosomal contigs predicted by plasmidEC are removed from the predictions. In the 'selection first' method, plasmid contigs are selected by plasmidEC and only these putative plasmid contigs are used as input for MOB-suite.

Plasmid predictions (bins) were aligned to their completed genome and analysed with regards to bin composition, recall and precision.

PlasmidEC successfully removes chromosomal contamination and doesn't affect recall of predictions

Standard MOB-suite output contained chromosomal contamination in 22.1% of bins, when using bactofidia as assembler, and in 24.5% of bins when assembly was performed with Unicycler. When applying plasmidEC with the 'reconstruction first' method, chromosomal contamination was reduced to 11.5% for both bactofidia and Unicycler assemblies. The 'selection first' method reduced contamination to 10.5% for bactofidia and 11.6% for Unicycler. However, this method produced fewer bins containing contigs derived from a unique plasmid (bactofidia: 418, Unicycler: 436) compared to the 'reconstruction first' method (bactofidia: 437, Unicycler: 464). Notably, the amount of bins composed of contigs that did not align to the reference genome was consistently higher in bactofidia assemblies, suggesting a higher rate of errors during assembly with this tool.
Using PlasmidEC to clean chromosomal contamination has no influence on the recall of bins and reference plasmids (Figure 10a and 10c). This result indicates that, in the majority of the cases, true plasmid contigs are not wrongfully removed from MOB-suite predictions. Additionally, by removing chromosomal contamination, plasmidEC increases precision of plasmid predictions (Figure 10b).

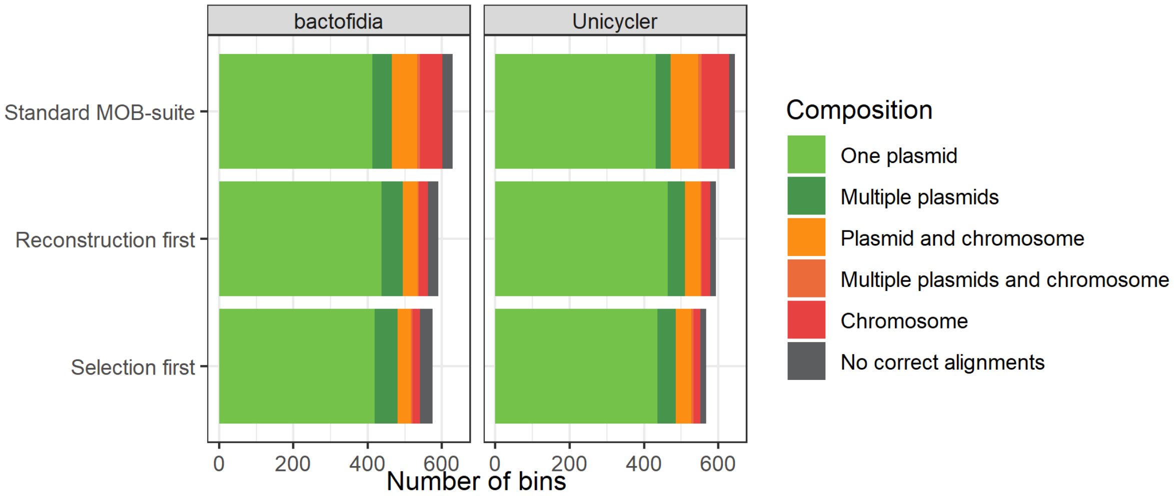


Figure 9 - Bin composition with respect to plasmid and chromosomal contigs for all bins predicted by MOB-suite. Showing results for standard MOB-suite and two methods that combine MOB-suite with plasmidEC ('reconstruction first' and 'selection first'). Input contigs were assembled by bactofidia and by Unicycler.

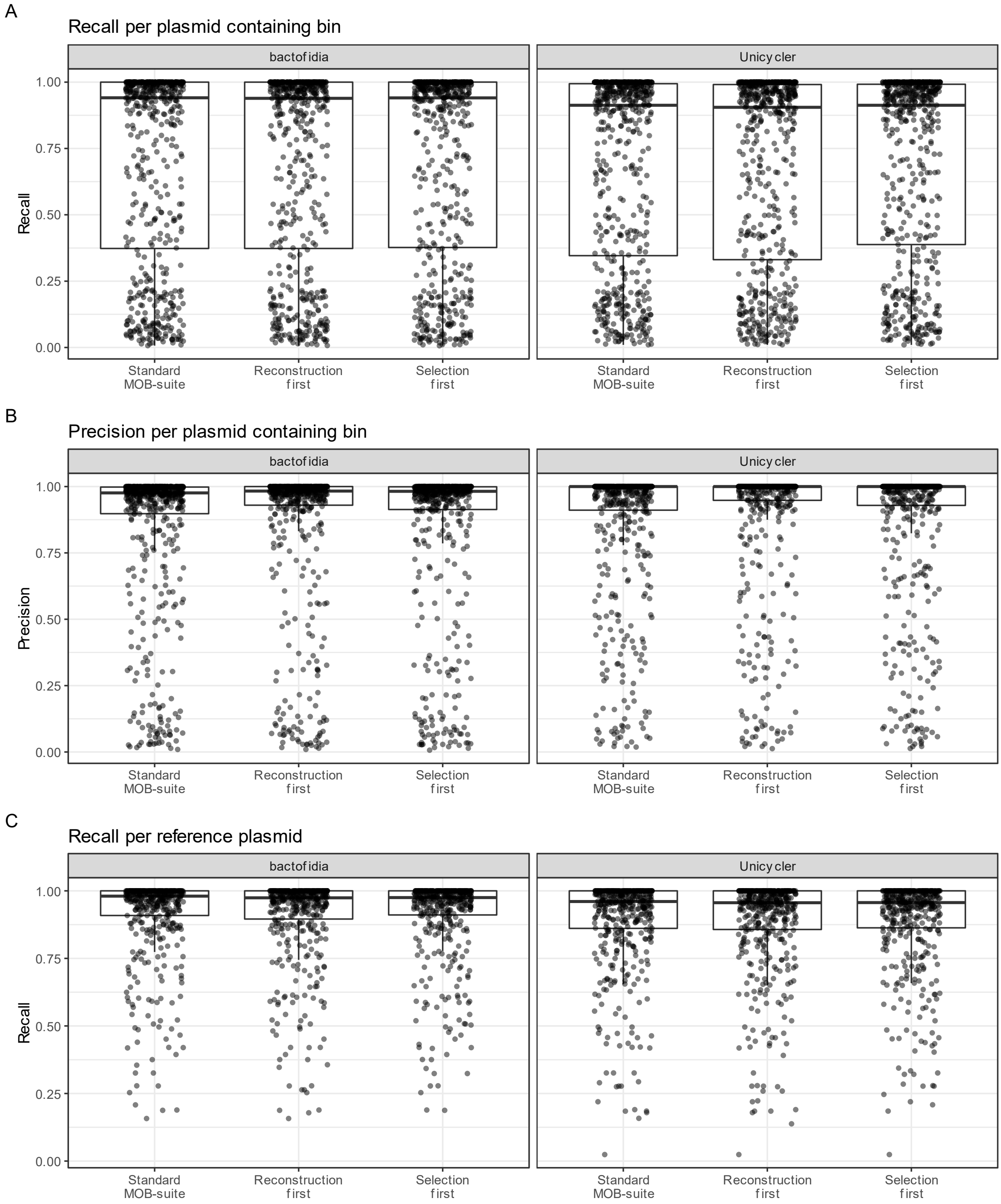


Figure 10 - Boxplots showing recall (a) and precision (b) for each bin. Also showing total recall per reference plasmid (c).

Discussion

In this work, we show that plasmidome identification from short-read sequencing data can be improved by combining the output of multiple binary classification tools. We developed plasmidEC, an ensemble classifier based on the majority vote between three input classifiers. The best software combination for *E. coli* was Platon/PlaScope/RFPlasmid, which presented an F1-score of 0.904 and outperformed all individual classifiers. We also demonstrated that plasmidEC identified a very large fraction of contigs derived from ARG-plasmids, scoring a recall of 0.941. In contrast, the recall of PlaScope, the best individual tool, was 0.884 for these plasmids. This means that ARG plasmid contigs that are missed by PlaScope can usually be recalled by Platon and RFPlasmid. Thus, plasmidEC is especially useful for plasmidome research that focuses on antibiotic resistance.
The differences in performance between PlaScope and plasmidEC for the entire benchmarking dataset are minor. Nevertheless, PlasmidEC shows a better balance between recall and precision. However, because multiple tools need to be run, plasmidEC requires more computational time and resources. All binary classifiers show improved recall for classifying contigs from ARG plasmids. It could be that these sequences are overrepresented in reference databases, which all tools use directly or indirectly.

There are many possibilities for further improvement of plasmidEC. Firstly, running predictions for multiple samples simultaneously should improve the speed of plasmidEC. Secondly, plasmidEC could be used to predict the plasmidome of species other than *E. coli*, as long as these are supported by the binary classifiers used as input. At the moment, PlaScope provides databases for *E. coli* and *Klebsiella*, but custom databases for other species can be created by the user.[21] Mlplasmids has models available for *E. coli*, *Enterococcus faecium*, *Enterococcus faecalis, Klebsiella pneumoniae* and *Acinetobacter baumannii*. Platon and RFPlasmid can already be used with any species. Of course, the performance of plasmidEC for species other than *E. coli* remains to be tested. It should be noted that the performance of the binary classifiers could decrease when making plasmidome predictions in species that are less frequently represented in reference databases.
Additional binary classifiers, whether novel or already existing, can be integrated into plasmidEC, which will allow to find better combinations between tools. Moreover, accuracy of plasmidEC could be improved by using weighted votes, where a prediction with higher confidence will count more heavily towards the final result than a low confidence prediction. A prerequisite is that input classifiers output a plasmid probability per contig instead of just a binary classification. Currently, such a probability score is only given by mlplasmids and RFPlasmid, but this could expand once new classifiers are added.

We show that plasmidEC can be used for plasmidome analysis of *E. coli* samples sequenced by short-read technology. We identified differences in the accessory genome, plasmidome and resistome between *E. coli* ST131 and other resistant *E. coli* isolates.
The pandemic pathogen *E. coli* ST131 shows a limited plasmidome and a conserved resistome. This suggests that limiting the size of the plasmidome and the diversity of ARGs could be beneficial for the overall success of this lineage. In fact, plasmids are known to inflict a substantial metabolic burden on their hosts, thereby reducing their fitness.[43] Furthermore, we found specific plasmidome-located ARG clusters in ST131, which are not commonly found in other *E. coli*. This appears to support the hypothesis that ST131 has developed stable relationships with specific ARG plasmids which are not easily disturbed and whose metabolic burden is alleviated by compensatory mutations.[44] In both ST131 and other STs we find mutually exclusive genes which encode a common resistance type. This mutual exclusion suggests that redundancy of function in resistance determinants is avoided by E. coli.
Additionally, ST131 shows an expansive accessory genome, which could be an indicator that this lineage possesses a larger repertoire of metabolic capabilities. Notably, the diversity in anaerobic metabolism pathways has been suggested as one of the keys for the success of *E. coli* ST131,

through enhancing host colonisation.[44–46] Functional analysis of the genes is necessary to validate this hypothesis.

Finally, we show that plasmidEC can be used to refine plasmid predictions of MOB-suite by removing chromosomal contamination. We tested this approach for contigs assembled by Unicycler and bactofidia, and evaluated whether the best results were achieved by running plasmidEC before MOB-suite (selection first) or running plasmidEC after MOB-suite (reconstruction first). Both methods were able to reduce chromosomal contamination to at least half the original value. The lowest amount of chromosomal contamination (10.5 %) was found using the 'selection first' method with bactofidia assembly. However, contigs assembled by bactofidia were more often unaligned to the reference genome than contigs assembled by Unicycler, indicating that bactofidia causes misassemblies more frequently. Unicycler is also the assembler recommended by the authors of MOB-suite.[35] We observe that the integration of plasmidEC with MOB-suite does not affect the recall of predictions, and improves precision. Thus, this is a valuable method to reduce chromosomal contamination in MOB-suite predictions.

Binary classifiers have been used in combination with plasmid reconstruction tools before. For example, in the initial step of gplas, plasmid contigs are detected using mlplasmids or PlasFlow. The choice of binary classifier has a major impact on the performance of gplas. Both recall and precision improve when using mlplasmids instead of PlasFlow. [49] We hypothesize that advancement in binary plasmid classification will consequently lead to improvements in plasmid reconstruction.

In conclusion, we show that our ensemble classifier plasmidEC successfully classifies *E. coli* contigs from short-read sequencing data as either plasmid- or chromosome-derived. Compared to existing binary classifiers, plasmidEC achieves increased recall, especially for contigs that derive from ARG plasmids. Plasmidome analysis using plasmidEC uncovered valuable differences between *E. coli* ST131 and other *E. coli*. We also show that plasmidEC refines individual plasmid reconstructions obtained with MOB-suite, by removing chromosomal contamination.

References

1. Cassini, A. *et al.* Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66 (2019).
2. Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report: 2021. https://www.who.int/publications-detail-redirect/9789240027336.
3. Leimbach, A., Hacker, J. & Dobrindt, U. E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity. in *Between Pathogenicity and Commensalism* (eds. Dobrindt, U., Hacker, J. H. & Svanborg, C.) 3–32 (Springer, 2013). doi:10.1007/82_2012_303.
4. Poirel, L. *et al.* Antimicrobial Resistance in Escherichia coli. *Microbiol. Spectr.* **6**, (2018).
5. Antimicrobial resistance in the EU/EEA (EARS-Net) - Annual Epidemiological Report for 2019. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019 (2020).
6. Bennett, P. M. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br. J. Pharmacol.* **153**, S347–S357 (2008).
7. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev. MMBR* **74**, 434–452 (2010).
8. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol.* **19**, 347–359 (2021).
9. Carattoli, A. Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* **303**, 298–304 (2013).
10. Levy, M. S., O'Kennedy, R. D., Ayazi-Shamlou, P. & Dunnill, P. Biochemical engineering approaches to the challenges of producing pure plasmid DNA. *Trends Biotechnol.* **18**, 296–305 (2000).
11. Smalla, K., Jechalke, S. & Top, E. M. Plasmid detection, characterization and ecology. *Microbiol. Spectr.* **3**, 10.1128/microbiolspec.PLAS-0038–2014 (2015).
12. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics* **3**, e000128 (2017).
13. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, (2017).
14. ten Doesschate, T. *et al.* In vivo acquisition of fosfomycin resistance in Escherichia coli by fosA transmission from commensal flora. *J. Antimicrob. Chemother.* **74**, 3630–3632 (2019).
15. Gan, H. M., Eng, W. W. H. & Dhanoa, A. First genomic insights into carbapenem-resistant Klebsiella pneumoniae from Malaysia. *J. Glob. Antimicrob. Resist.* **20**, 153–159 (2020).
16. Van Driessche, L. *et al.* Isolation of Drug-Resistant Gallibacterium anatis from Calves with Unresponsive Bronchopneumonia, Belgium. *Emerg. Infect. Dis.* **26**, 721–730 (2020).
17. Gupta, S. K., Shin, H., Han, D., Hur, H.-G. & Unno, T. Metagenomic analysis reveals the prevalence and persistence of antibiotic- and heavy metal-resistance genes in wastewater treatment plant. *J. Microbiol.* **56**, 408–415 (2018).
18. Arredondo-Alonso, S. *et al.* Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen Enterococcus faecium. *mBio* **11**, (2020).
19. Schwengers, O. *et al.* Platon: identification and characterization of bacterial plasmid contigs in short--read draft assemblies exploiting protein sequence--based replicon distribution scores. *Microb. Genomics* 12 (2020).
20. Arredondo-Alonso, S. *et al.* mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genomics* **4**, (2018).
21. Royer, G. *et al.* PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb. Genomics* **4**, (2018).
22. Bloois, L. van der G. van, Wagenaar, J. A. & Zomer, A. L. RFPlasmid: Predicting plasmid sequences from short read assembly data using machine learning. *bioRxiv* 2020.07.31.230631 (2020) doi:10.1101/2020.07.31.230631.
23. Paganini, J. A., Plantinga, N. L., Arredondo-Alonso, S., Willems, R. J. L. & Schürch, A. C. Recovering Escherichia coli Plasmids in the Absence of Long-Read Sequencing Data. *Microorganisms* **9**, 1613 (2021).
24. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021).
25. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies.

Bioinforma. Oxf. Engl. **29**, 1072–1075 (2013).

26. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).
27. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. 2018. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb. Genomics* **4**, e000192.
28. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
29. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
30. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
31. Meyer, D. & Buchta, C. *proxy: Distance and Similarity Measures*. (2021).
32. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
33. Griffith, D. M., Veech, J. A. & Marsh, C. J. cooccur: Probabilistic Species Co-Occurrence Analysis in R. *J. Stat. Softw. Code Snippets* **69**, 1–17 (2016).
34. Almende B.V. and Contributors, Thieurmel, B. & Robert, T. *visNetwork: Network Visualization using 'vis.js' Library*. (2021).
35. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genomics* **4**, (2018).
36. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
37. Manges, A. R. *et al.* Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. *Clin. Microbiol. Rev.* **32**, (2019).
38. Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.* **27**, 1437–1449 (2017).
39. Wittekamp, B. H. *et al.* Decontamination Strategies and Bloodstream Infections With Antibiotic-Resistant Microorganisms in Ventilated Patients: A Randomized Clinical Trial. *JAMA* **320**, 2087–2098 (2018).
40. Petty, N. K. *et al.* Global dissemination of a multidrug resistant Escherichia coli clone. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5694–5699 (2014).
41. Harris, P. N. A. *et al.* Whole genome analysis of cephalosporin-resistant Escherichia coli from bloodstream infections in Australia, New Zealand and Singapore: high prevalence of CMY-2 producers and ST131 carrying blaCTX-M-15 and blaCTX-M-27. *J. Antimicrob. Chemother.* **73**, 634–642 (2018).
42. Schürch, A. C. bactofidia. *GitLab* https://gitlab.com/aschuerch/bactofidia.
43. San Millan, A. & MacLean, R. C. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiol. Spectr.* **5**, (2017).
44. Cummins, E. A., Snaith, A. E., McNally, A. & Hall, R. J. The role of potentiating mutations in the evolution of pandemic Escherichia coli clones. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* (2021) doi:10.1007/s10096-021-04359-3.
45. Bonnet, R. *et al.* Host Colonization as a Major Evolutionary Force Favoring the Diversity and the Emergence of the Worldwide Multidrug-Resistant Escherichia coli ST131. *mBio* **12**, e0145121 (2021).
46. McNally, A. *et al.* Diversification of Colonization Factors in a Multidrug-Resistant Escherichia coli Lineage Evolving under Negative Frequency-Dependent Selection. *mBio* **10**, e00644-19 (2019).
47. Nicolas-Chanoine, M.-H., Bertrand, X. & Madec, J.-Y. Escherichia coli ST131, an Intriguing Clonal Group. *Clin. Microbiol. Rev.* **27**, 543–574 (2014).
48. Kondratyeva, K., Salmon-Divon, M. & Navon-Venezia, S. Meta-analysis of Pandemic Escherichia coli ST131 Plasmidome Proves Restricted Plasmid-clade Associations. *Sci. Rep.* **10**, 36 (2020).
49. Arredondo-Alonso, S. *et al.* gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics* **36**, 3874–3876 (2020).

Supplementary figures

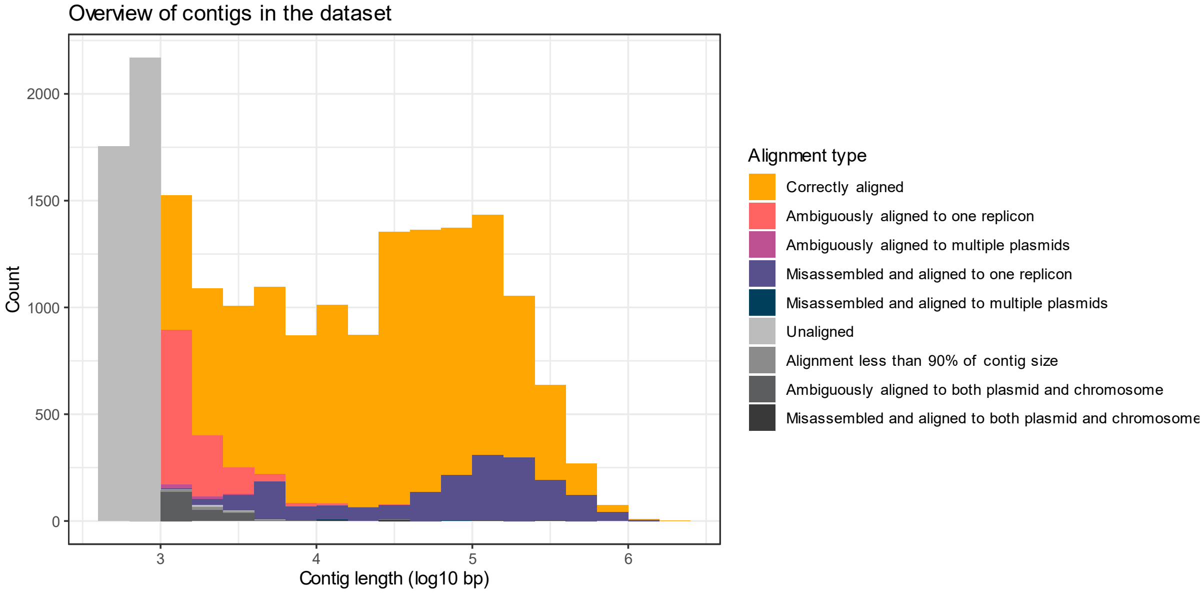


Figure S1 - Alignment types of all contigs in the dataset by contig length. Included contigs are shown in colour, excluded contigs in greyscale.

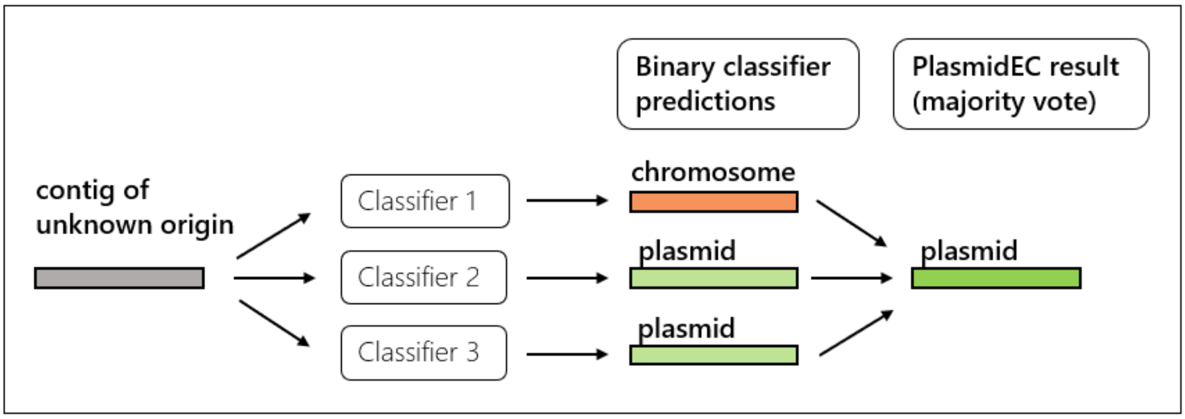


Figure S2 - Overview of the majority voting system method applied by plasmidEC.

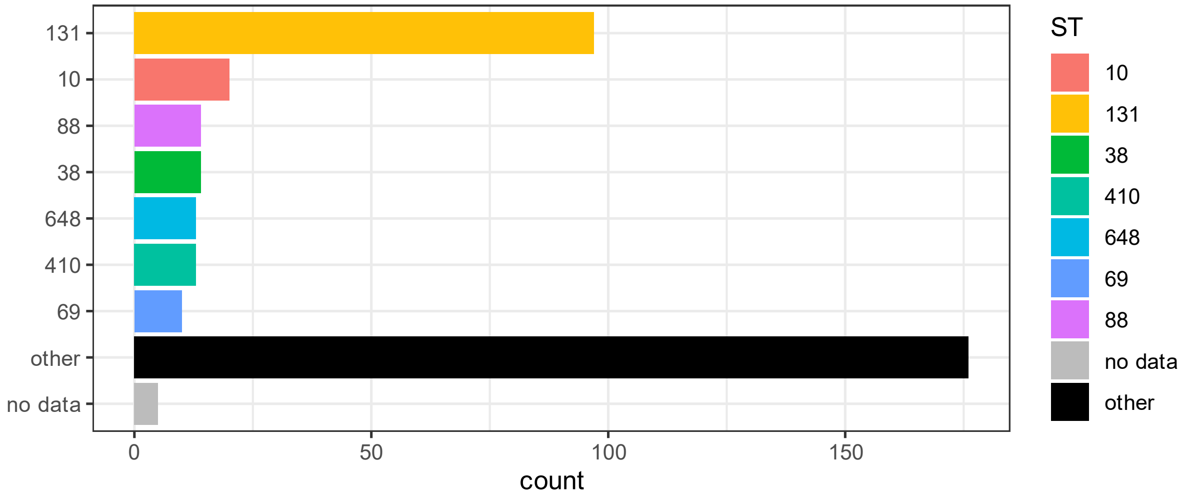


Figure S3 - Prevalence of sequence types (STs) in the dataset. Only counting STs present in at least 10 samples.

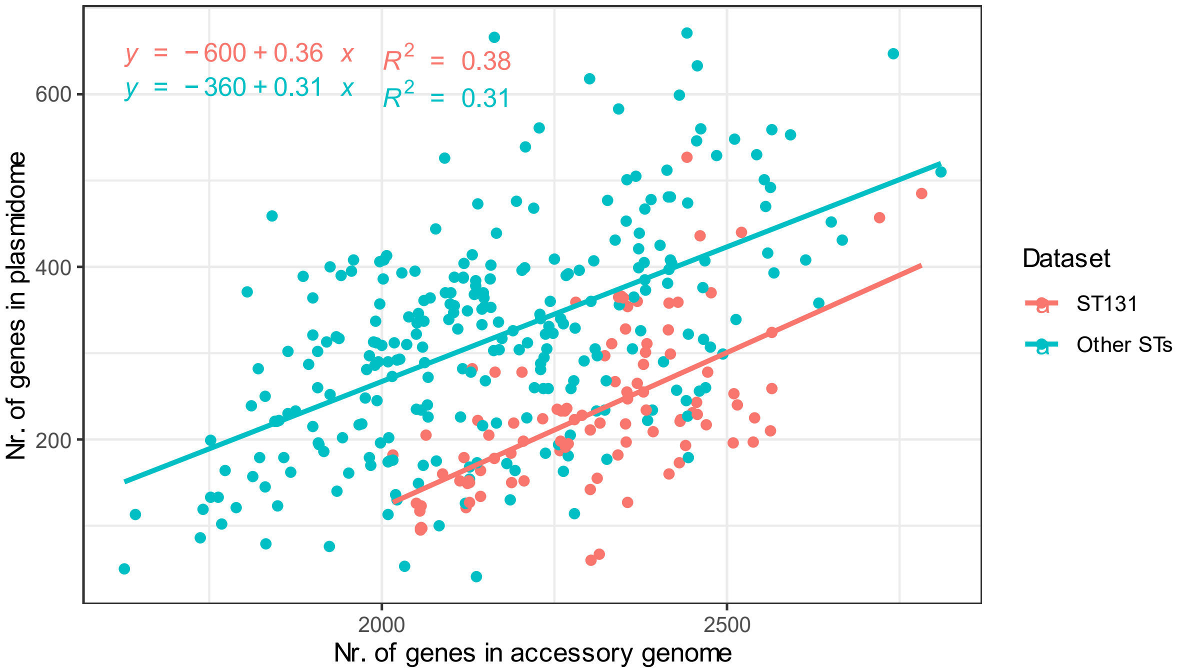


Figure S4 - Correlation between plasmidome size and accessory genome size for *E. coli* ST131 (red) and other STs (blue). Showing formula of linear regression and r squared value.

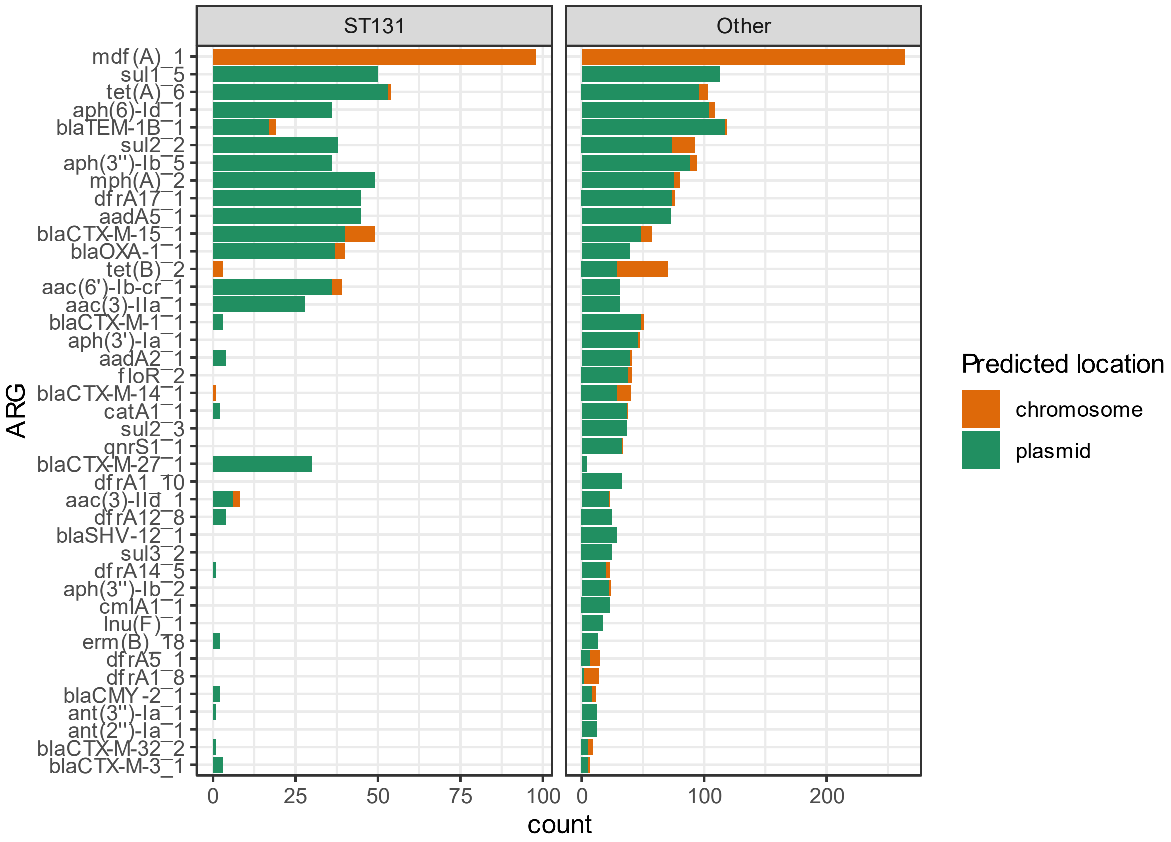


Figure S5 - Prevalence of ARGs in *E. coli* ST131 and other STs, ordered by their prevalence in the full dataset. ARGs are coloured by their predicted location; plasmid (green) or chromosome (orange). Only showing ARGs present in at least 10 samples.

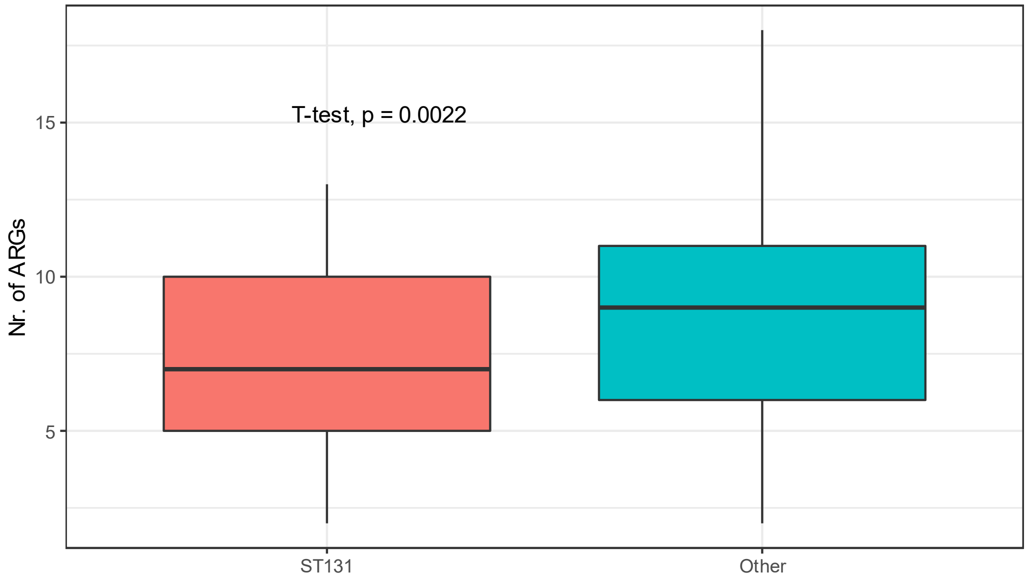
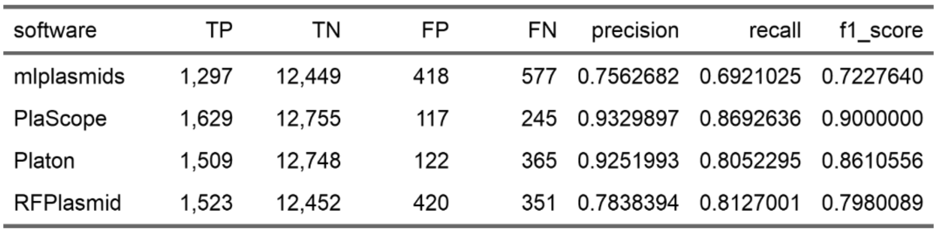


Figure S6 - Number of genes present in the resistome of *E. coli* ST131 (red) compared to other STs (blue). Significance determined by Welch t-test.

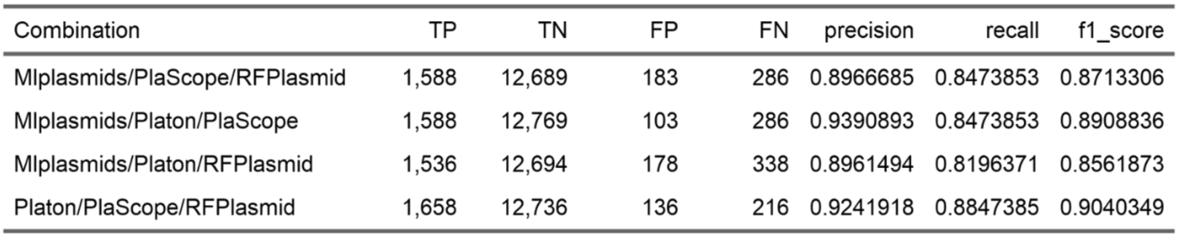
Supplementary tables

Table S1 - True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), precision, recall and F1-score for the evaluated binary classifiers



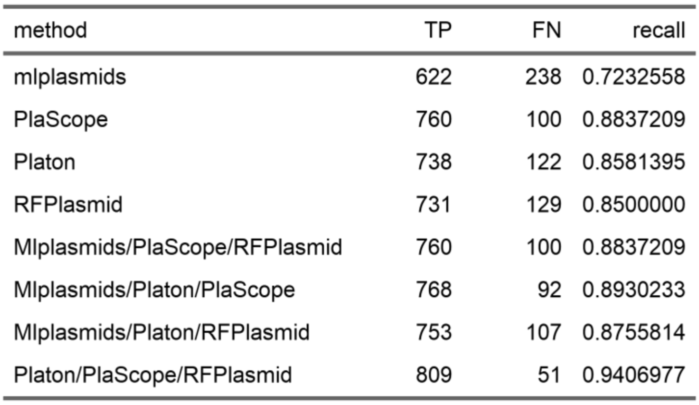
software	TP	TN	FP	FN	precision	recall	f1_score
mlplasmids	1,297	12,449	418	577	0.7562682	0.6921025	0.7227640
PlaScope	1,629	12,755	117	245	0.9329897	0.8692636	0.9000000
Platon	1,509	12,748	122	365	0.9251993	0.8052295	0.8610556
RFPlasmid	1,523	12,452	420	351	0.7838394	0.8127001	0.7980089

Table S2 - True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), precision, recall and F1-score for all combinations of plasmidEC



Combination	TP	TN	FP	FN	precision	recall	f1_score
Mlplasmids/PlaScope/RFPlasmid	1,588	12,689	183	286	0.8966685	0.8473853	0.8713306
Mlplasmids/Platon/PlaScope	1,588	12,769	103	286	0.9390893	0.8473853	0.8908836
Mlplasmids/Platon/RFPlasmid	1,536	12,694	178	338	0.8961494	0.8196371	0.8561873
Platon/PlaScope/RFPlasmid	1,658	12,736	136	216	0.9241918	0.8847385	0.9040349

Table S3 - True positives (TP), false negatives (FN) and recall of binary classifiers and plasmidEC combinations evaluated for ARG-plasmids (antibiotic resistance gene containing plasmids).



method	TP	FN	recall
mlplasmids	622	238	0.7232558
PlaScope	760	100	0.8837209
Platon	738	122	0.8581395
RFPlasmid	731	129	0.8500000
Mlplasmids/PlaScope/RFPlasmid	760	100	0.8837209
Mlplasmids/Platon/PlaScope	768	92	0.8930233
Mlplasmids/Platon/RFPlasmid	753	107	0.8755814
Platon/PlaScope/RFPlasmid	809	51	0.9406977