

# PlasmidEC and gplas2: an optimized short-read approach to predict and reconstruct antibiotic resistance plasmids in *Escherichia coli*

Julian A. Paganini<sup>1</sup>, Jesse J. Kerkvliet<sup>1</sup>, Lisa Vader<sup>1</sup>, Nienke L. Plantinga<sup>1</sup>, Rodrigo Meneses<sup>1</sup>, Jukka Corander<sup>2,3,4</sup>, Rob J. L. Willems<sup>1</sup>, Sergio Arredondo-Alonso<sup>2,3,†</sup> and Anita C. Schürch<sup>1,\*†</sup>

## Abstract

Accurate reconstruction of *Escherichia coli* antibiotic resistance gene (ARG) plasmids from Illumina sequencing data has proven to be a challenge with current bioinformatic tools. In this work, we present an improved method to reconstruct *E. coli* plasmids using short reads. We developed plasmidEC, an ensemble classifier that identifies plasmid-derived contigs by combining the output of three different binary classification tools. We showed that plasmidEC is especially suited to classify contigs derived from ARG plasmids with a high recall of 0.941. Additionally, we optimized gplas, a graph-based tool that bins plasmid-predicted contigs into distinct plasmid predictions. Gplas2 is more effective at recovering plasmids with large sequencing coverage variations and can be combined with the output of any binary classifier. The combination of plasmidEC with gplas2 showed a high completeness (median=0.818) and F1-Score (median=0.812) when reconstructing ARG plasmids and exceeded the binning capacity of the reference-based method MOB-suite. In the absence of long-read data, our method offers an excellent alternative to reconstruct ARG plasmids in *E. coli*.

## DATA SUMMARY

No new sequencing data have been generated in this study. All genomes used in this research are publicly available at GenBank and Sequence Read Archive of the National Center for Biotechnology Information. Accession numbers are specified in the Supplementary Materials. Scripts to reproduce the results reported in this paper can be accessed at <https://gitlab.com/jpaganini/ecoli-binary-classifier>. The ensemble classifier, plasmidEC, is publicly available at <https://gitlab.com/mmb-umcu/plasmidEC/-/releases/1.4.0>, and gplas2 (release 1.0.0) can be found at <https://gitlab.com/mmb-umcu/gplas2/-/releases/1.0.0>.

## INTRODUCTION

*Escherichia coli* is a commensal Gram-negative bacterium inhabiting the gastrointestinal tract but is also the leading cause of bloodstream and urinary tract infections in humans [1, 2]. In recent years, the emergence and spread of multidrug-resistant *E. coli* lineages has limited the treatment options for such infections [3, 4]. Moreover, a recent assessment of the global burden of antimicrobial resistance (AMR) estimated that AMR *E. coli* infections accounted for more than 250000 deaths in 2019, placing *E. coli* as one of the most prevalent AMR pathogens worldwide [5].

Horizontal gene transfer is one of the main drivers behind the rapid spread of AMR [6–8]. Antibiotic resistance genes (ARGs) are commonly associated with mobile genetic elements (MGEs), which facilitate their mobility across bacteria [9, 10]. Out of these

Received 19 September 2023; Accepted 22 January 2024; Published 20 February 2024

**Author affiliations:** <sup>1</sup>Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands; <sup>2</sup>Department of Biostatistics, Faculty of Medicine, University of Oslo, Oslo, Norway; <sup>3</sup>Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK; <sup>4</sup>Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.

**\*Correspondence:** Anita C. Schürch, a.c.schurch@umcutrecht.nl

**Keywords:** antibiotic resistance; assembly graph; bioinformatics; *Escherichia coli*; Illumina; plasmids; short reads; WGS.

**Abbreviations:** AMR, antimicrobial resistance; ARG, antibiotic resistance gene; ESBL, extended-spectrum beta lactamase; FN, false negatives; FP, False Positives; IQR, interquartile range; IS, insertion sequence; ST, sequence type; WGS, whole genome sequencing.

†These authors contributed equally to this work

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

001193 © 2024 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

**Impact Statement**

*Escherichia coli* has emerged as a highly pervasive multidrug-resistant pathogen on a global scale. The dissemination of resistance is significantly influenced by plasmids, mobile genetic elements that facilitate the transfer of antimicrobial resistance genes within and between diverse bacterial species. Consequently, precise and high-throughput identification of plasmids is imperative for effective genomic surveillance of resistance. However, accurate plasmid reconstruction remains challenging with the use of affordable short-read sequencing data. In this work, we present a novel method to accurately predict and reconstruct *E. coli* plasmids based on Illumina data. Additionally, we demonstrate that our approach outperforms the reference-based method MOB-suite, especially when reconstructing plasmids carrying antimicrobial resistance genes.

MGEs, plasmids play a pivotal role by disseminating AMR in clinical settings as well as in other environments [11–13]. Plasmids are frequently transmitted among bacteria of the same species, but they can also be shared between bacteria of different species or even different genera [14–17]. Given their relevance in the spread of AMR genes, it is critical to develop high-throughput methods to identify plasmids in a precise, fast and accessible manner.

Bacterial genomes have been massively studied using short-read sequencing platforms. However, plasmids tend to contain repetitive elements that cannot be spanned by short-reads and thus their sequence is usually fragmented into several contigs and mingled with other genomic elements. This makes it hard to reconstruct complete plasmids from short-read sequencing data [18].

Several fully automated bioinformatic tools are currently available to predict plasmids from short-read sequencing data. They can be broadly categorized into two groups: (i) tools that produce a binary classification of contigs as either plasmid- or chromosome-derived, predicting the total plasmid content of a bacterial strain, often referred to as the ‘plasmidome’ (without reconstructing individual plasmids), and (ii) tools that aim to recover complete sequences for individual plasmids [19]. The latter group, termed plasmid reconstruction tools, provides a more suitable output for plasmid epidemiology studies.

We recently evaluated the performance of several plasmid reconstruction tools for use with *E. coli* short-read data [19]. We found that the best performing tool, MOB-suite [20], only achieved the correct reconstruction in 50.2% of the plasmids. Moreover, all tools underperformed when attempting to reconstruct plasmids containing antibiotic resistance genes (ARG-plasmids), ranging from 3.4 to 27.9% correct ARG-plasmid reconstructions. These results emphasized the need to improve current methods to predict ARG-plasmids in *E. coli*.

Here, we present a new high-throughput method to reconstruct *E. coli* plasmids from short-read sequencing data. First, we optimized gplas [21], a plasmid binning tool, to compute walks in the assembly graph corresponding to plasmids with a pronounced coverage variation. Second, we developed an ensemble classifier, plasmidEC, combining multiple existing binary classification tools (Centrifuge [22] coupled to PlaScope’s database [23], RFplasmid [24], Platon [25] and mlplasmids [26]) to predict plasmid-derived contigs. Coupling plasmidEC with gplas2 allowed us to accurately bin plasmid-derived contigs into separate components corresponding to individual plasmid sequences. Our method outperforms all plasmid reconstruction tools previously evaluated in Paganini *et al.* [19], especially for predicting ARG-plasmids.

**METHODS**

All scripts used to reproduce the analyses can be found at [gitlab.com/jpaganini/ecoli-binary-classifier](https://gitlab.com/jpaganini/ecoli-binary-classifier). R version 3.6.1. was used for all R scripts.

**Benchmark datasets**

A dataset of 240 complete *E. coli* genomes from eight different phylogroups and 117 sequence types (STs), carrying 631 plasmids, was selected as previously described in Paganini *et al.* [19]. Samples were isolated from animals, humans and the environment, resulting in a diverse dataset with respect to phylogeny and plasmid content. All genome sequences were completed by the combination of short- and long-read sequencing data. Short-read sequences and complete genomes were downloaded from NCBI using SRA tools (v2.10.9) and ncbi-genome-download (v0.2.10) (<https://github.com/kblin/ncbi-genome-download>), respectively. Genomes present in the training datasets or reference databases of existing plasmid classification tools (mlplasmids, Centrifuge, Platon and/or RFPlasmid) were removed ( $n=26$ ). The remaining 214 samples, carrying 542 plasmids, were used to benchmark the binary classifiers (Data S1, available in the online version of this article). From these, 15 genomes (Data S2) were randomly selected for optimization of the gplas algorithm and excluded from later comparisons. The remaining genomes ( $n=199$ , 483 plasmids) were used to benchmark the plasmid reconstruction methods.

## Benchmarking binary classification tools and construction of plasmidEC

### Selection of contigs for benchmarking

Short-read sequences of each sample were assembled with *bactofidia* (v1.1) (<https://gitlab.com/aschuerch/bactofidia>), a pipeline that relies on SPAdes for genome assembly (v3.11.1) [27]. The resulting contigs ( $n=18\,963$ ) were labelled as chromosome- or plasmid-derived by alignment to their respective complete genomes using QUAST (v5.0.2) [28]. Only contigs larger than 1000 bp with an alignment of at least 90% the contig length were considered ( $n=15\,020$ ). Of those, contigs aligning to multiple positions in the genome (ambiguously aligned contigs) were included as long as they exclusively aligned to either the chromosome or to plasmids ( $n=1236$ ). The same criterion was used for the inclusion of misassembled contigs ( $n=1862$ ). In total, the benchmark dataset included 14746 contigs (Fig. S1). SPAdes was selected as an assembler to retain equivalence of our methods with those employed in other benchmark experiments, as referenced previously [23–26].

### Binary classification of contigs using Centrifuge

Centrifuge is a tool that serves as a taxonomy classifier for metagenomics reads. We adapted this tool to function as a binary classifier of whole genome sequencing (WGS) bacterial contigs, similarly to what has been described for PlaScope [23]. The database developed for PlaScope was used for classification. In contrast to PlaScope, our Centrifuge-based classifier does not filter contigs based on size or coverage.

### Assessment of the individual binary classifiers

Contigs were classified by *mlplasmids* (v2.1.20), *Centrifuge* (v1.0.4\_beta), *Platon* (v1.7) and *RFPlasmid* (v0.0.18). All tools were run using default parameters. We assessed the performance of the four binary classifiers by comparing, for each contig, their prediction to the true class of the contig, as described in the section above. For Centrifuge, an ‘unclassified’ prediction was handled as a negative prediction. Predictions were categorized 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). Global performance of the tools was evaluated with the following metrics:

$$\begin{aligned} \text{Recall}(\text{contig}) &= \frac{TP}{TP+FN} \\ \text{Precision}(\text{contig}) &= \frac{TP}{TP+FP} \\ \text{F1 Score}(\text{contig}) &= 2 \cdot \frac{\text{Recall}(\text{contig}) \cdot \text{Precision}(\text{contig})}{\text{Recall}(\text{contig}) + \text{Precision}(\text{contig})} \end{aligned}$$

### Assessment of the ensemble classifiers

To improve the predictions obtained by independent tools, we combined their output into distinct ensemble classifiers that implemented a majority voting system. We tested four different combinations of individual classifiers: *mlplasmids*/*Centrifuge*/*Platon*, *mlplasmids*/*Centrifuge*/*RFPlasmid*, *mlplasmids*/*Platon*/*RFPlasmid* and *Centrifuge*/*Platon*/*RFPlasmid*. A final classification of each contig (chromosome or plasmids) was obtained by combining the output of the tools using an R script (provided in the accompanying code repository). The ensemble classifiers were evaluated using the same metrics as described above.

### Construction of plasmidEC

The tool consists of a bash wrapper script that automatically installs and runs all required individual classifiers and combines their results with a majority voting system. Based on the performance for *E. coli*, the combination of *Centrifuge*/*Platon*/*RFPlasmid* was selected as the default. *PlasmidEC* is publicly available at <https://gitlab.com/mmb-umcu/plasmidEC>.

## Benchmarking plasmid reconstruction tools

### Running plasmid predictions tools

Prior to assembly, Illumina raw reads were trimmed using *trim-galore* (v0.6.6) (<https://github.com/FelixKrueger/TrimGalore>) to remove bases with a Phred quality score below 20. *Unicycler* (v0.4.8) [29], explicitly recommended by the authors of *MOB-suite* for optimal performance [20], was then applied to perform *de novo* assembly with default parameters. Contigs larger than 1000 bp were used as input for *MOB-suite* (v3.0.0) [20], while assembly graphs in GFA format served as input for *gplas2* (v2.0.0). To run *gplas2*, nodes from the graph were first classified as plasmid- or chromosome-derived using either *plasmidEC* or *Centrifuge*; only nodes larger than 1000 bp were classified. Output from the tools was modified to assign probabilities for the classification of each node, which is required by the *gplas2* algorithm. For *Centrifuge*, discrete probabilities were assigned based on the node classification status; if a node was classified as plasmid, a probability of 1 was assigned, while chromosome-predicted nodes were assigned zero. In the case of unclassified nodes, a probability of 0.5 was assigned. By default, *plasmidEC* assigns probabilities based on the fraction of tools that agreed on the classification. For example,

if two out of three tools agreed on classifying a node as plasmid, a probability of 0.66 is assigned. Plasmid reconstruction results shown as 'gplas\_mplasmids' were obtained from our previous benchmark study [19].

### Analysis of the plasmid bin composition

To evaluate the bins created by MOB-suite, gplas and gplas2, we used QUAST (v5.0.2) [28] to align the contigs of each bin to the respective complete reference genome. We calculated accuracy, completeness and F1-score on the base-pair level, as specified below:

$$\text{Accuracy (bp)} = \frac{\text{Alignment length against reference plasmid (bp)}}{\text{Total length of predicted bin (bp)}}$$

$$\text{Completeness (bp)} = \frac{\text{Alignment length against reference plasmid (bp)}}{\text{Total length of reference plasmid (bp)}}$$

$$\text{F1 score (bp)} = 2 \cdot \frac{\text{Accuracy (bp)} \cdot \text{Completeness (bp)}}{\text{Accuracy (bp)} + \text{Completeness (bp)}}$$

If a bin was composed of contigs derived from different plasmids, then accuracy<sub>(bp)</sub>, completeness<sub>(bp)</sub> and F1-Score<sub>(bp)</sub> were reported for each plasmid-bin combination.

We also evaluated the number of reference plasmids that were detected by each tool. We consider a reference plasmid as detected when at least a single contig of the plasmid was included in the predictions.

To determine *combined completeness* for each reference plasmid, all bins generated in an isolate were combined as follows:

$$\text{Combined completeness (bp)} = \sum_{i=1}^n \text{Completeness (bp)}$$

$n$  = Total number of bins that contain contigs aligning to the reference plasmid.

where  $n$  is the total number of bins that contain contigs aligning with the reference plasmid.

### Antibiotic resistance gene prediction

Resistance genes were predicted by running Abricate (v1.0.1) against the Resfinder [30] database (database indexed on 19 April 2020) with reference plasmids as query, using 80% as the identity and coverage cut-off. The same software and parameters were used to predict the presence of ARGs in the plasmid-predicted contig bins generated by each of the plasmid reconstruction tools.

### Evaluation of ARG binning

For bins that carried ARGs, we calculated recall<sub>(ARG)</sub> and precision<sub>(ARG)</sub> as indicated below:

$$\text{Recall (ARG)} = \frac{\text{Nr. of correctly predicted ARGs in bin}}{\text{Total nr. of ARGs in reference plasmid}}$$

$$\text{Precision (ARG)} = \frac{\text{Nr. of correctly predicted ARGs in bin}}{\text{Total nr. of ARGs in bin}}$$

### Evaluating unbinned nodes in gplas predictions

Unitigs classified as unbinned by gplas ( $n=78$ ) were aligned to the corresponding complete reference genome using QUAST (v5.0.2). The results of these alignments were used to determine the origin of the unitig (plasmid or chromosome). For isolates that contained more than one unbinned unitig ( $n=19$ ), coverage information of all unitigs (bin and unbinned) was extracted from the header of the FASTA files generated after unicycler assembly. From these data, coverage variance for all replicons was calculated and plotted using R (v.3.6.1).

### Evaluating the recovered fraction for each reference plasmid

We calculated the maximum completeness<sub>(bp)</sub> that can be obtained to reconstruct every reference plasmid using short-read sequencing data. Before applying any classification tool, all nodes from the assembly graph were converted to FASTA format using the 'extract' option of gplas2. Nodes smaller than 1000 bp or smaller than 500 bp were filtered out using seqtk (v1.3) (<https://github.com/lh3/seqtk>), and remaining nodes were aligned to their respective complete reference genomes using QUAST to obtain the completeness<sub>(bp)</sub> values. The completeness<sub>(bp)</sub> value was called the *recovered fraction*.

### Read coverage of missing reference plasmids

A small number of plasmids were either completely missed or recovered with low completeness after short-read assembly. To determine if these sequences were also missing from short-reads, trimmed Illumina reads were aligned to reference genomes using BWA MEM (v0.7.17) [31] with default parameters. The resulting SAM files were converted to BAM and sorted using SAMtools (v1.9) [32]. Read coverages per base were determined using BEDTOOLS (v2.30.0) [33].

## RESULTS

### Optimization of gplas to improve the reconstruction of *E. coli* plasmids

The gplas algorithm performs *de novo* reconstruction of plasmids through multiple steps (Fig. 1, Steps 1 to 3) [21]. In short, nodes from the assembly graph are initially classified as plasmid-derived or chromosome-derived by an external binary classification software, which also assigns a probability to the classifications. Then, plasmid-predicted unitigs act as seeds to compute plasmid walks with homogeneous coverage in the assembly graph using a greedy approach. Finally, these unitigs are binned together into individual components based on their coexistence in the computed plasmid walks. A detailed description of the algorithm can be found in the original publication [21]. Given that gplas performed sub-optimally when reconstructing *E. coli* plasmids in our previous study [19], in gplas2 we introduced two major modifications to the algorithm.

#### A. Expansion of the input options for binary classification

Coupling gplas with an accurate binary classifier improves the reconstruction of plasmids, as we have previously demonstrated for *Enterococcus faecalis* and *Klebsiella pneumoniae* [21, 34]. Consequently, the gplas2 algorithm accepts predictions from any binary classifier, provided they output classification probabilities and expected file formats.

#### B. Re-iterating plasmid walks over initially unbinned contigs

Gplas constructs plasmid walks over the assembly graph to connect unitigs that potentially originate from the same plasmid (Fig. 1, Step 2). Consequently, plasmid-predicted unitigs that cannot be connected to other unitigs through these walks are classified as unbinned, and are not included in the plasmid predictions (Fig. 1, Step 3). Unbinned unitigs seem to originate from reference plasmids that were sequenced with a pronounced coverage variation (Fig. S2). This sequencing artefact poses a challenge to the gplas algorithm, which builds plasmid walks from unitigs with homogeneous coverage. Consequently, we modified gplas to consider these coverage variations (Fig. 1, Steps 4 and 5). Whenever unbinned unitigs are produced, gplas2 will generate a second round of binning in bold mode by running two additional steps.

#### Computation of plasmid walks in bold mode starting from unbinned unitigs

If unbinned unitigs are predicted, new bold plasmid walks will be constructed. When creating the bold walks, a higher coverage variance threshold between plasmid-predicted unitigs is allowed. This threshold can be defined by the user and is a multiple of the coverage variance observed for chromosome-predicted unitigs. Only bold plasmid walks that start from unbinned unitigs will be retained to use in the next step, while the rest will be discarded (Fig. 1, Step 4).

#### Plasmidome network reconstruction and repartitioning

Plasmid walks produced during bold mode are merged with plasmid walks from normal mode. Based on these combined data, plasmidome networks are reconstructed and repartitioned (Fig. 1, Step 5) to create new bins, using the same algorithms as in Step 3.

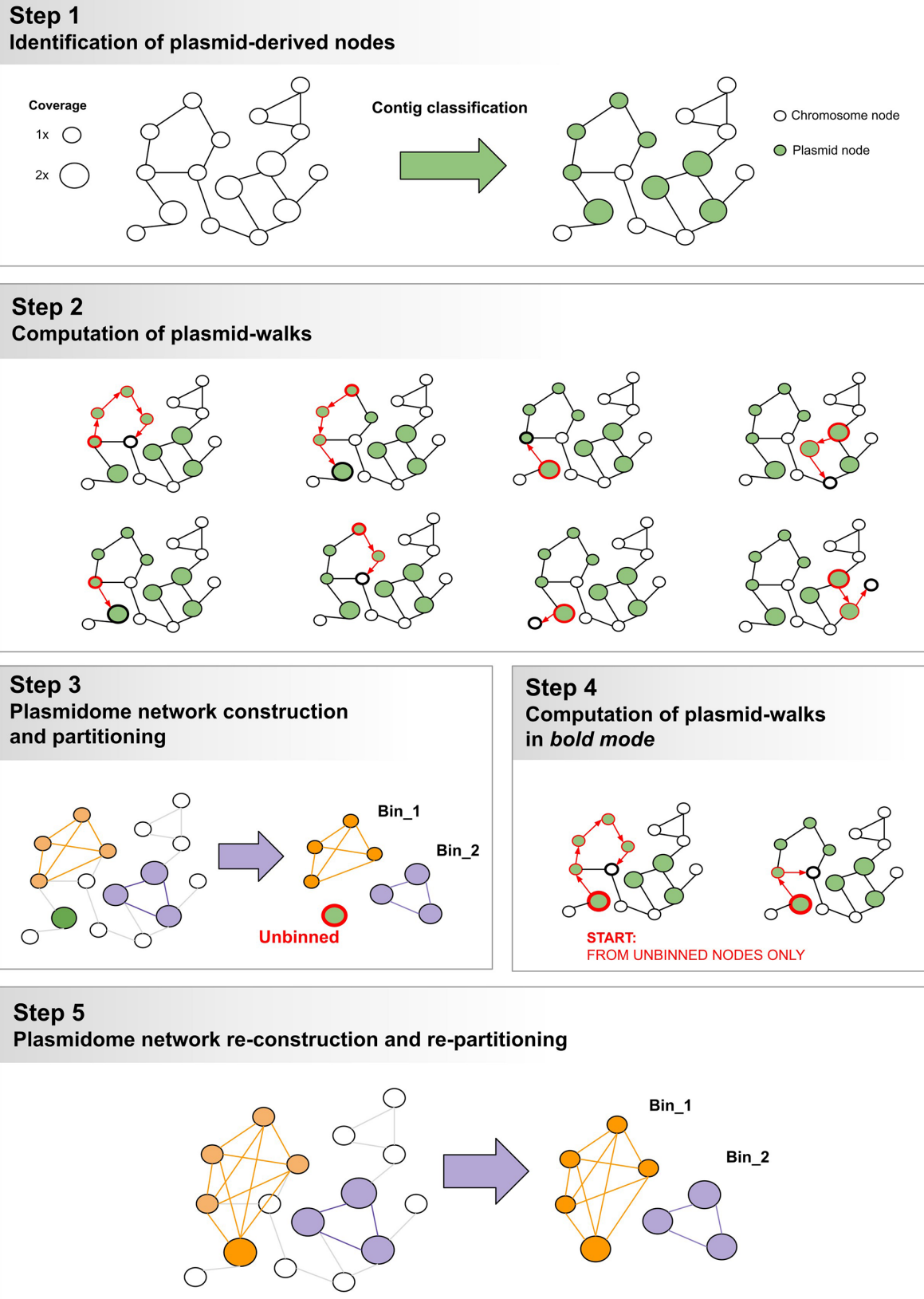
We optimized the predictions obtained with gplas2 using a subset of 15 *E. coli* genomes that contained unbinned unitigs and that were excluded from subsequent benchmarking efforts (Data S2). For bold walks, we allowed a coverage variance of 5, 10, 15 or 20 times the coverage variance observed for the chromosome-predicted unitigs. Plasmid predictions made with gplas2 exhibited consistently higher completeness<sub>(bp)</sub> values when compared to the original predictions (Fig. S3A). Surprisingly, altering the coverage variance threshold above 5 did not impact completeness<sub>(bp)</sub> values. In contrast, accuracy<sub>(bp)</sub> values decreased when allowing a higher coverage variance. The highest F1-Score<sub>(bp)</sub> values [median=0.78, interquartile range (IQR)=0.47–0.96] were obtained when using a coverage variance threshold of 5. Consequently, 5 was defined as the default value to construct bold plasmid walks. As a single example, we display the plasmid predictions obtained with and without running bold mode for genome GCA\_013823335.1\_ASM1382333v1 (Fig. S3B and S3C). In this case, the bold walks allowed recovery of seven additional contigs belonging to plasmids CP057179.1 and CP057180.1.

Gplas2, including the aforementioned features and a detailed user guide, can be found at <https://gitlab.com/mmb-umcu/gplas2>.

### Comparing binary classification methods for *E. coli*

In order to combine gplas2 with the best available binary classifier for *E. coli*, we compared the performance of four different tools (Centrifuge, RFPlasmid, mlplasmids and Platon). The benchmark dataset consisted of 14746 contigs. Of these contigs, 87.3% ( $n=12\,872$ ) were chromosome-derived and 12.7% ( $n=1874$ ) were plasmid-derived, as determined by alignment to complete reference genomes.

We evaluated the number of contigs which were correctly and incorrectly classified by each of the tools and calculated recall<sub>(contig)</sub>, precision<sub>(contig)</sub> and F1-Score<sub>(contig)</sub> (Table S1). Centrifuge was able to correctly identify the highest number of plasmid-derived contigs (TP  $n=1629$ ), while the rest of the tools detected between 1297 and 1523 plasmid-derived contigs. Notably, Centrifuge also included the least chromosomal contamination in its predictions (FP,  $n=117$ ), closely followed by Platon ( $n=122$ ). In contrast, mlplasmids and RFPlasmid included a higher amount of chromosome-derived contigs in their plasmidome predictions



**Fig. 1.** Schematics of the gplas2 algorithm. Steps 4 and 5 were added to gplas2 in order to recover unbinned unitigs.

( $n=418$  and  $n=420$ , respectively). Centrifuge was the tool with the highest F1-Score<sub>(contig)</sub> (0.900) followed by Platon (0.861), RFPlasmids (0.798) and mlplasmids (0.722). For most tools, precision<sub>(contig)</sub> values were higher than recall<sub>(contig)</sub> values, indicating that the predicted plasmidome mostly consists of true plasmid-derived contigs, but also that plasmid contigs were frequently missed by the tools.

We also explored the congruence in contig classifications across tools (Fig. 2). All tools agreed on the correct classification of 51.8% of plasmid-derived contigs (TP:  $n=971$ , Fig. 2a), and another 26.5% of plasmid-derived contigs were correctly classified by at least three tools ( $n=497$ ). Also, a large fraction (94.1%) of chromosome-derived contigs were correctly classified by all tools (TN:  $n=12116$ , Fig. 2b). Moreover, only a minority of plasmid-derived and chromosome-derived contigs were missed by most of the tools and correctly classified by just a single tool (TP: 85/1874, 4.7%, TN: 58/12,872, 0.5% respectively). From these observations, we concluded that contig misclassifications are primarily derived from individual tools (Fig. 2c, d).

### PlasmidEC: a voting classifier for improved detection of ARG-plasmid contigs in *E. coli*

We theorized that discarding software-specific misclassifications, while keeping correct classifications shared by multiple tools, could improve the overall binary classification of *E. coli* contigs as plasmid- or chromosome-derived. To explore this, we combined the predictions of three individual classifiers and extracted their majority vote as the final classification.

After testing all possible combinations of individual classifiers, we found that Platon/Centrifuge/RFPlasmid displayed the highest overall performance of voting classifiers with the highest F1-Score<sub>(contig)</sub> (0.904). This ensemble classifier achieved an F1-Score<sub>(contig)</sub> similar to Centrifuge (0.900) but had a slightly higher recall<sub>(contig)</sub> (0.884 and 0.869, respectively) (Fig. 3a, b, Table S1).

Next, we evaluated recall<sub>(contig)</sub> values for a subset of plasmids ( $n=114$ ) encoding ARGs (ARG-plasmids) (Fig. 3c, d, Table S2). This dataset consisted of 860 plasmid-derived contigs, derived from 91 *E. coli* genomes. The recall<sub>(contig)</sub> of individual tools ranged from 0.723 (mlplasmids) to 0.884 (Centrifuge), whereas the different combinations of tools in a voting classifier reached recall<sub>(contig)</sub> values ranging from 0.883 (mlplasmids/Platon/RFPlasmid) to 0.941 (Platon/Centrifuge/RFPlasmid).

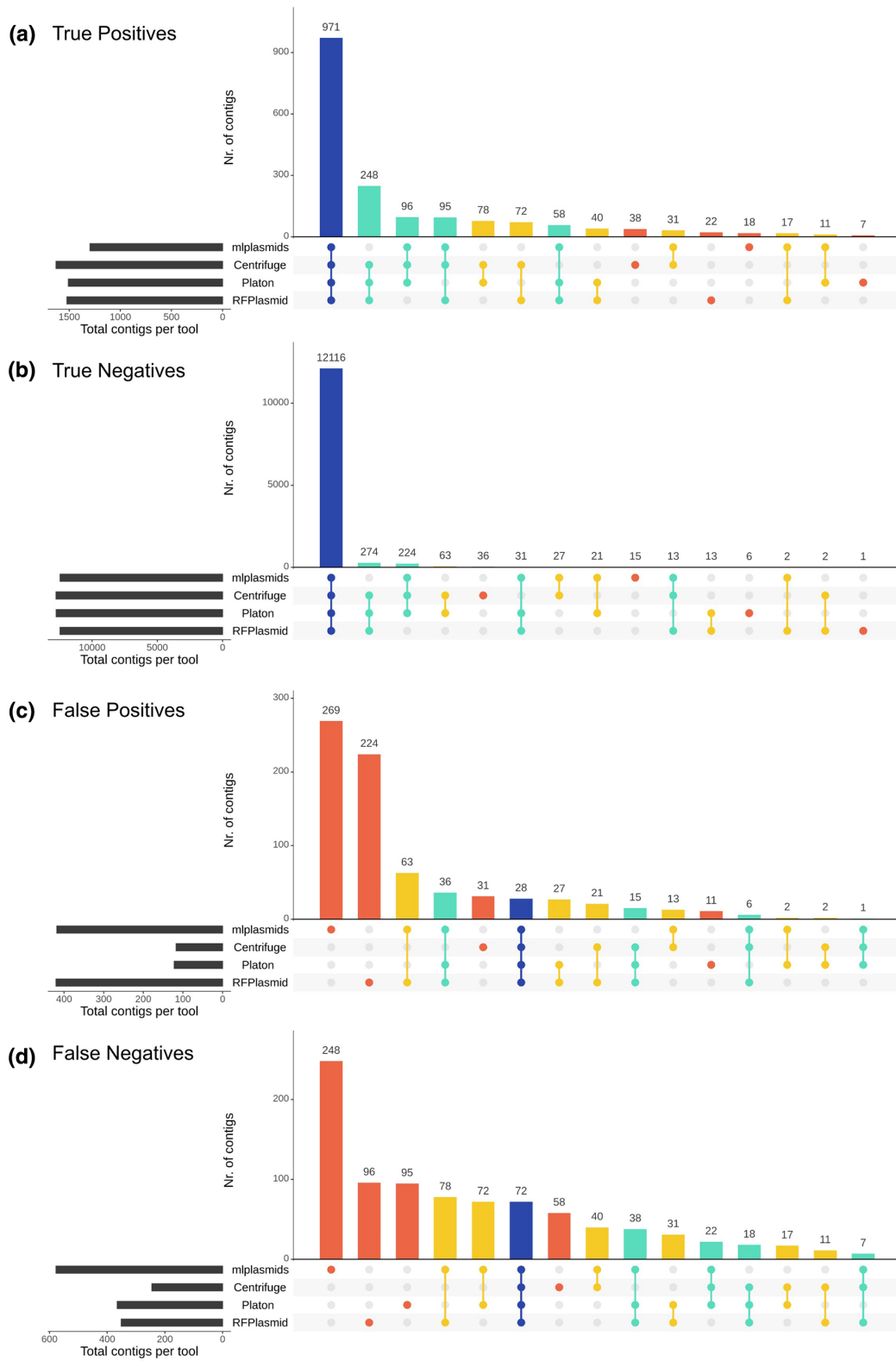
Based on these results, the combination of Platon/PlaScope/RFPlasmid was selected as the ensemble classifier to be implemented in a novel tool termed plasmidEC, which is publicly available at <https://gitlab.com/mmb-umcu/plasmidEC>.

We measured the computational resources used by the ensemble and individual classifiers (Fig. S4). Binary classifiers showed considerable differences in both CPU time and memory usage. The average CPU time required per sample was lowest for Centrifuge (0.2 min) and highest for Platon (14.9 min). 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/Centrifuge/RFPlasmid required the least number of resources (CPU time=4.5 min, memory=9.0 Mb) and Centrifuge/Platon/RFPlasmid the most (CPU time=21.5 mins, memory=21.4 Mb).

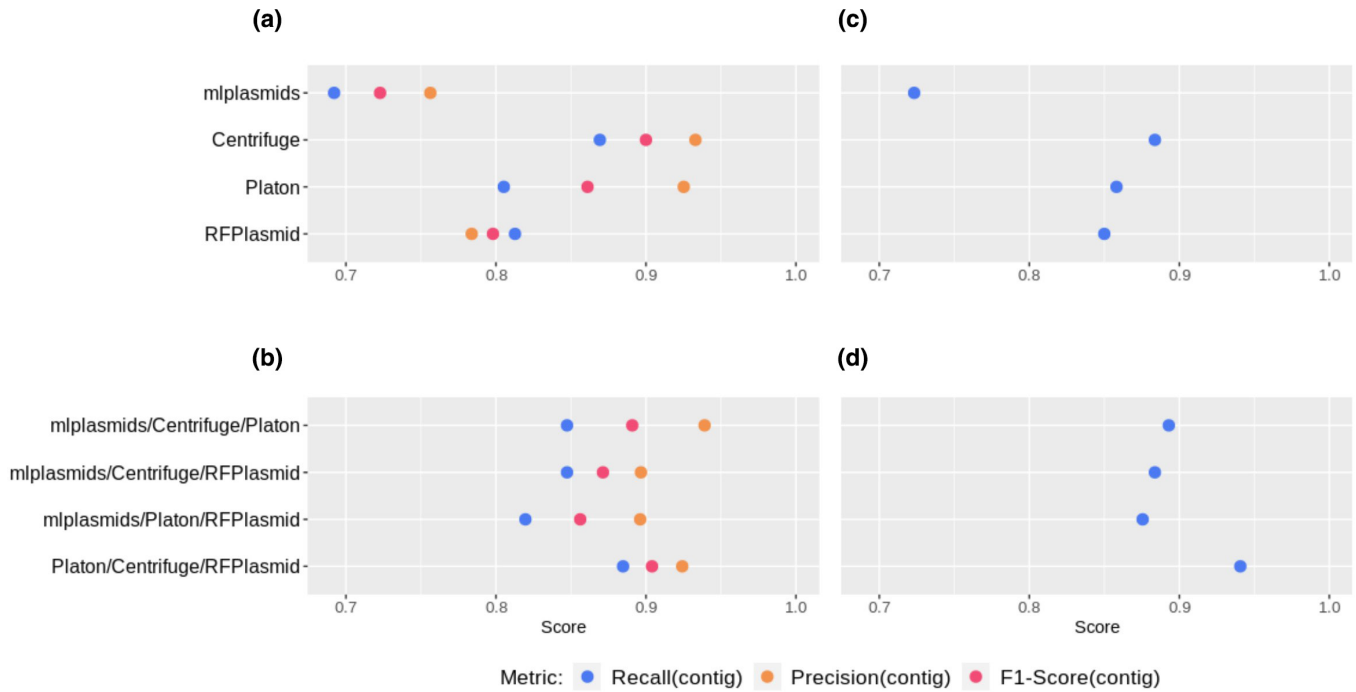
### Exploiting the information from the assembly graph improves correct binning of ARG plasmids

To reconstruct individual *E. coli* plasmids, gplas2 was combined with plasmidEC and Centrifuge, and performance was compared against a previous version of gplas (which relied on mlplasmids for binary classification of contigs) and against MOB-suite, which was the best-performing plasmid reconstruction tool for *E. coli* in our recent benchmark study [19]. To retain comparability with the aforementioned study, we started with the same dataset and removed 26 genomes that were present in the Centrifuge database and 15 genomes that were used to improve the gplas2 algorithm. Consequently, our benchmark dataset consisted of 199 complete *E. coli* genomes, which carried 483 plasmids. A total of 213 (44.1%) plasmids were classified as small (<18 000 bp), while the remaining 270 (55.9%) were large [19]. Given our interest in predicting ARG-plasmids, and the fact that most ARGs are encoded on large plasmids ( $n=382/387$ , 98.7%), we analysed performance separately for large ARG-plasmids ( $n=96$ ) and large non-ARG-plasmids ( $n=174$ ).

When evaluating the reconstruction of ARG-plasmids, we found that the F1-Score<sub>(bp)</sub> values of gplas2 combined with either plasmidEC (gplas2\_plasmidEC) or Centrifuge (gplas2\_Centrifuge) were similar (Fig. 4a, Table 1). However, gplas2\_plasmidEC (median=0.81, IQR=0.53–0.93) performed slightly better than gplas2\_Centrifuge (median=0.76, IQR=0.52–0.94). Notably, both gplas2 methods outperformed MOB-suite (median=0.44, IQR=0.18–0.87) and gplas\_mlplasmids (median=0.46, IQR=0.24–0.68). As accuracy<sub>(bp)</sub> values were nearly identical across tools, the disparity in F1-Scores<sub>(bp)</sub> can be explained due to the differences in completeness<sub>(bp)</sub>. Gplas\_mlplasmids displayed diminished combined completeness<sub>(bp)</sub> values, indicating a restricted ability to detect plasmid-derived contigs, probably caused by the low recall observed in mlplasmids (Fig. 3c). In contrast, combined completeness<sub>(bp)</sub> distributions were virtually identical among gplas2 and MOB-suite, thus suggesting that these methods had a similar capacity to detect contigs derived from ARG-plasmids, but gplas2 performed better at binning these contigs together into individual predictions. This hypothesis was confirmed by analysing the number of bins into which each reference plasmid was fragmented (Fig. 4b). For ARG-plasmids, we found that MOB-suite fragmented 49% of plasmids into multiple predictions, while both gplas2 methods did so in only 14% of cases.



**Fig. 2.** Upset diagrams showing congruence in contig classification by different binary prediction tools (absolute counts). 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). Bar colours indicate the number of tools that concur in the classification of the contigs.



**Fig. 3.** Performance of individual binary classifiers and plasmidEC combinations, measured by recall<sub>(contig)</sub>, precision<sub>(contig)</sub> and F1-Score<sub>(contig)</sub>. (a) Individual classifiers evaluated using the full dataset ( $n=214$  genomes). (b) PlasmidEC combinations evaluated using the full dataset. (c) Individual classifiers evaluated using a dataset of ARG-plasmids ( $n=114$  plasmids). (d) PlasmidEC combinations evaluated using a dataset of ARG-plasmids.

Next, we evaluated the capacity of the tools to detect plasmid-derived ARGs (Fig. 4c, Table 1). MOB-suite, gplas2\_plasmidEC and gplas2\_Centrifuge performed similarly, detecting 336 (86.8%), 336 (86.8%) and 332 (85.8%) ARGs, respectively. Moreover, these tools successfully detected all ARGs present in small plasmids ( $n=5$ , 100%). In contrast, gplas\_mplasmids only detected 221 (55.5%) plasmid-borne ARGs and failed to detect any ARGs present in small plasmids. In concordance with previous results, recall<sub>(ARG)</sub> values (Fig. 4d) for gplas2 predictions were higher than those obtained with MOB-suite and gplas\_mplasmids (Table 1). This indicates that gplas2 performs better at correctly binning ARGs together into the same bin. However, plasmid predictions made with gplas2 also included a higher number of chromosome-derived ARGs (Fig. 4c, Table 1).

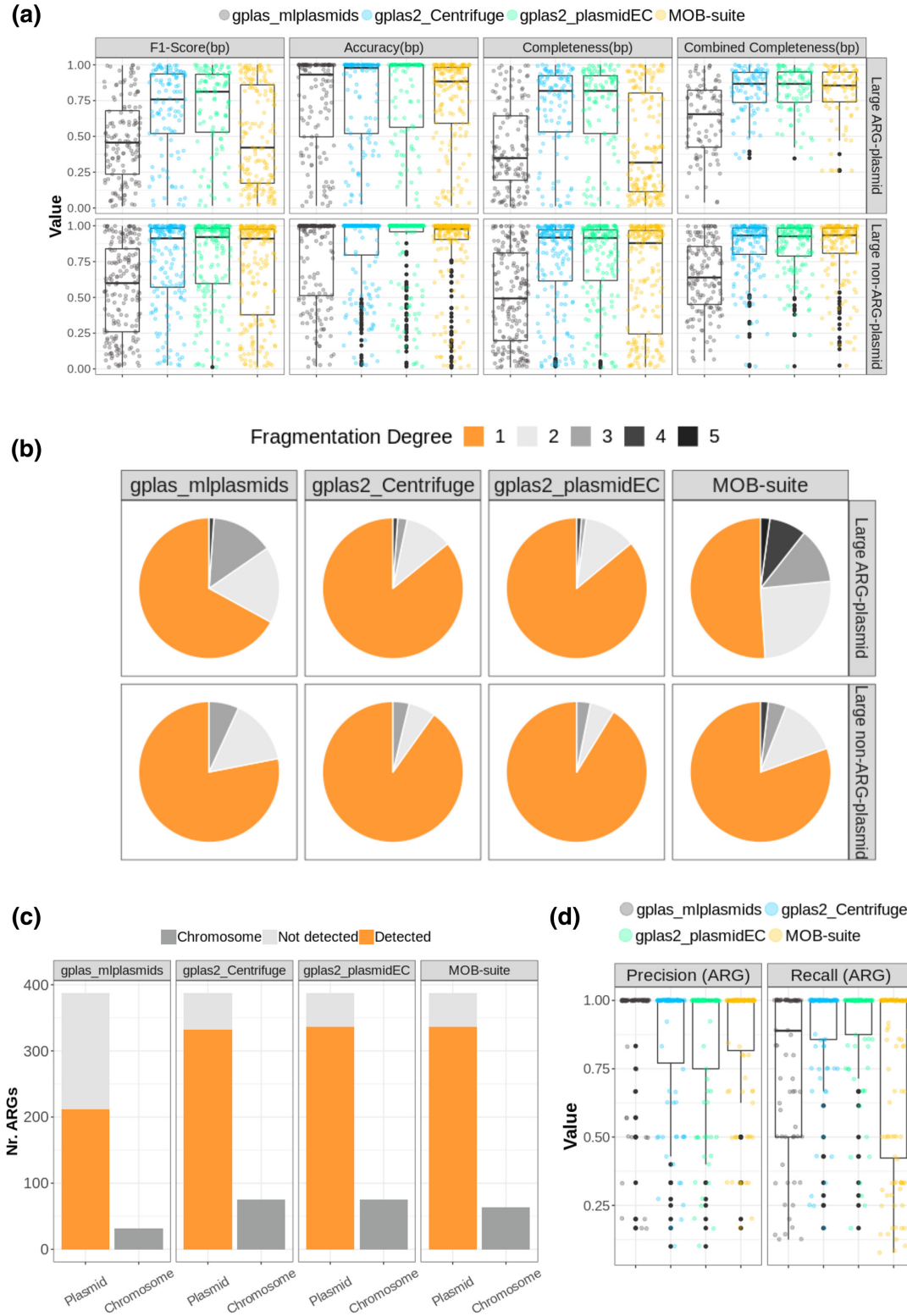
Interestingly, gplas2 methods and MOB-suite performed similarly well when reconstructing extended spectrum beta-lactamase (ESBL) plasmids ( $n=42$ ). MOB-suite reconstructions were characterized by having higher accuracy<sub>(bp)</sub> and gplas2 methods reconstructed ESBL-plasmids with higher completeness<sub>(bp)</sub> (Fig. S5A). Despite the aforementioned differences, these three tools exhibited similar F1-Score<sub>(bp)</sub> values. Additionally, the number of plasmid-borne ESBL genes detected was almost identical across tools (Fig. S5B). In contrast, gplas\_mplasmids presented lower F1-Score<sub>(bp)</sub> values and detected fewer plasmid-borne ESBL genes. Out of all tools, gplas2 methods performed slightly better at binning ARGs into the same prediction (Fig. S5C).

For small plasmids ( $n=213$ ), all tools displayed similar performance across the three metrics, obtaining near-perfect reconstructions in all cases, with F1-Score<sub>(bp)</sub> medians of 1 (Fig. S6A, Table 1). This is probably due to most small plasmids being assembled into a single contig ( $n=196$ , 92.0%) (Fig. S6B), and consequently the identification of these contigs as plasmid-derived generally leads to obtaining high values for all metrics. We therefore evaluated the number of small (and large) plasmids detected by each of the tools (Fig. S6C, Table 1). Interestingly, gplas2\_Centrifuge detected 196 (92.0%) small plasmids, and gplas2\_plasmidEC performed similarly, detecting 184 (86.4%). Both gplas2 methods outperformed MOB-suite and gplas\_mplasmids, which detected 174 (81.79%) and 82 (38.5%) small plasmids, respectively.

Finally, we tested the effect of using different contig size cut-offs for plasmid reconstruction. We found no significant differences in performance of the tools when using 500 or 1000 bp as the minimum contig size. A more detailed description of the results from this analysis can be found in the Supplementary Materials and in Figs S7–S10.

## DISCUSSION

Accurately reconstructing *E. coli* plasmids from Illumina reads has proven to be a challenge, especially in the context of ARG-plasmids. In this work, we developed a new high-throughput method to reconstruct *E. coli* plasmids *de novo* from short-read



**Fig. 4.** Benchmarking of plasmid reconstruction methods. (a) Completeness<sub>(bp)</sub>, accuracy<sub>(bp)</sub>, F1-Score<sub>(bp)</sub> and combined completeness<sub>(bp)</sub> values for predictions corresponding to large ARG-plasmids ( $n=96$ ) and large non-ARG-plasmids ( $n=174$ ). (b) Percentage of reference plasmids that were recovered with different fragmentation degrees (i.e. if contigs belonging to a reference plasmid are assigned to three different predictions, then the fragmentation degree equals three). (c) Absolute count of ARGs included (detected) in plasmid predictions, missing ARGs (not detected) and chromosome-derived ARGs incorrectly included (Chromosome). (d) Recall<sub>(ARG)</sub> and precision<sub>(ARG)</sub> values.

**Table 1.** Performance summary of three plasmid prediction tools, for the prediction of different plasmid types

	MOB-suite	gplas2_plasmidEC	gplas2_Centrifuge	gplas_mlplasmids
<b>Large plasmids (n=270)</b>				
No. of detected plasmids*	263 (97.4%)	253 (93.7%)	254 (94.1%)	237 (87.8%)
<b>ARG-plasmids (n=96)</b>				
F1-Score <sub>(bp)</sub> (median, IQR)	0.421 (0.172–0.860)	0.812 (0.529–0.934)	0.758 (0.520–0.936)	0.457 (0.236–0.680)
Completeness <sub>(bp)</sub> (median, IQR)	0.317 (0.114–0.803)	0.818 (0.520–0.924)	0.818 (0.531–0.924)	0.349 (0.194–0.643)
Accuracy <sub>(bp)</sub> (median, IQR)	0.883 (0.591–0.982)	0.979 (0.564–1)	0.979 (0.520–1)	0.932 (0.497–1)
No. of plasmid-borne ARGs detected	336 (86.8%)	336 (86.8%)	332 (85.8%)	212 (55.5%)
No. of chromosome-derived ARGs	64	75	75	32
Recall <sub>(ARG)</sub> (median, IQR)	1 (0.42–1)	1 (0.86–1)	1 (0.86–1)	0.89 (0.50–1)
Precision <sub>(ARG)</sub> (median, IQR)	1 (0.82–1)	1 (0.75–1)	1 (0.77–1)	1 (1–1)
<b>Non-ARG-plasmids (n=174)</b>				
F1-Score <sub>(bp)</sub> (median, IQR)	0.910 (0.378–0.977)	0.921 (0.596–0.983)	0.912 (0.571–0.983)	0.600 (0.260–0.840)
Completeness <sub>(bp)</sub> (median, IQR)	0.879 (0.245–0.967)	0.915 (0.618–0.972)	0.918 (0.614–0.972)	0.493 (0.198–0.810)
Accuracy <sub>(bp)</sub> (median, IQR)	0.978 (0.904–1)	1 (0.958–1)	1 (0.796–1)	0.994 (0.513–1)
<b>Small plasmids (n=213)</b>				
No. of detected plasmids*	174 (81.8%)	184 (86.4%)	196 (92.0%)	82 (38.5%)
F1-Score <sub>(bp)</sub> (median, IQR)	1 (0.985–1)	1 (0.991–1)	1 (0.990–1)	0.972 (0.948–0.980)
Completeness <sub>(bp)</sub> (median, IQR)	1 (0.976–1)	1 (0.996–1)	1 (0.990–1)	0.981 (0.953–0.986)
Accuracy <sub>(bp)</sub> (median, IQR)	1 (1–1)	1 (1–1)	1 (1–1)	0.968 (0.946–0.978)
No. of plasmid-borne ARGs detected	5 (100%)	5 (100%)	5 (100%)	0 (0%)

\*A plasmid is considered detected if at least one contig is included in the plasmid predictions.

sequencing data. Our method relies on an accurate identification of plasmid-derived nodes in the assembly graph, followed by the binning of these nodes using sequencing coverage and node connectivity information. We proved that our method outperforms other plasmid prediction tools available for *E. coli*, especially when reconstructing ARG-plasmids.

To improve the identification of plasmid-derived contigs, we built plasmidEC, an ensemble classifier that combines predictions from three individual binary classifiers and implements a majority voting system. Voting classifiers have been successfully applied in other fields of biology [35–38], but so far not for the problem of plasmidome identification. PlasmidEC correctly identified a large fraction of contigs derived from ARG-plasmids (recall<sub>(contig)</sub>=0.941), and considerably outperformed all individual classifiers. Thus, we believe that plasmidEC will be especially useful for plasmidome research that focuses on antibiotic resistance. Notably, all binary classifiers presented higher recall<sub>(contig)</sub> for classifying contigs from ARG plasmids than from non-ARG plasmids, suggesting that these sequences might be overrepresented in reference databases which are directly or indirectly used by all tools.

When comparing the performance of the tools using the entire benchmark dataset, we found that plasmidEC and Centrifuge performed very similarly in terms of F1-Score<sub>(contig)</sub>. However, plasmidEC showed a higher recall<sub>(contig)</sub> but used more computational resources and took a longer time to complete the predictions. Reference-based methods, such as Centrifuge, are expected to perform well for species like *E. coli* which are abundant in public databases [39]. Supporting this hypothesis, a recent study by Shaw et al. [40] discovered very few novel plasmid sequences in a dataset that included more than 2000 plasmids from *Enterobacteriaceae* isolates. Centrifuge [22] is a metagenomics classifier designed to predict the origin of sequences based on custom databases. Recently, it was also shown that the usage of Kraken [41], another metagenomic classifier using customized databases, outperformed other binary classifiers in *Klebsiella pneumoniae* [22, 42]. It would be interesting to explore how tools perform at classifying contigs from species with a limited number of complete genomes in databases. We speculate that in those cases, plasmidEC, which combines tools with diverse computational approaches, could improve predictions to a larger extent.

PlasmidEC could be further optimized by (i) multithreading the predictions of the individual tools, which would reduce the computational time to generate the results, (ii) including the possibility to predict the origin of contigs from other species, as long

as those are supported by the binary classifiers, and (iii) improving its accuracy by using weighted votes, where a high-confidence prediction will contribute more to the final result than a low-confidence prediction.

We integrated plasmidEC (and Centrifuge) with gplas2 to reconstruct individual *E. coli* plasmids. We then compared the performance of gplas2 combined with those classifiers against both MOB-suite and a previous iteration of gplas, which uses mlplasmids as a binary classifier. Interestingly, the most pronounced differences in performance were observed when reconstructing ARG-plasmids. Although combined completeness<sub>(bp)</sub> values indicated that gplas2 methods and MOB-suite identified similar fractions of ARG-plasmids, MOB-suite more frequently fragmented ARG-plasmids into multiple bins, yielding low completeness<sub>(bp)</sub> and F1-Score<sub>(bp)</sub>. In contrast, gplas2 (either with plasmidEC or Centrifuge) was more successful at binning together contigs into individual plasmid predictions, thus achieving considerably higher values for the aforementioned metrics. The previous version of gplas presented lower values of combined completeness<sub>(bp)</sub>, probably due to the constrained ability of mlplasmids to identify contigs originating from ARG-plasmids. Accuracy<sub>(bp)</sub> values for all tools were very similar, indicating a similar degree of chimeric predictions. Interestingly, both gplas2 methods performed similarly to MOB-suite when reconstructing plasmids that carry ESBL genes, which suggests that these plasmids might be overrepresented in the database used by MOB-suite to make predictions.

We recently described that ARG-plasmids from *E. coli* are particularly difficult to reconstruct from short-read data [19], and we suggested that the modular nature of these plasmids could complicate their reconstruction using strict reference-based methods, such as MOB-suite. The results we obtained here seem to confirm this hypothesis. Additionally, we improved the reconstruction of ARG-plasmids by using coverage and node connectivity information. Yet, our study also proves that enriching the assembly graph with accurate information on the origin of contigs (plasmid/chromosome) is equally important. A previous version of gplas, which used mlplasmids as a binary classifier, performed significantly worse at predicting ARG-plasmids in *E. coli* [19]. Moreover, using a simpler graph-based approach that mainly relies on coverage differences to identify plasmids is also insufficient. This approach, applied by plasmidSPAdes, frequently leads to the inclusion of chromosomal contamination [18, 19], due to the low copy number that ARG-plasmids often exhibit.

We envision that gplas2 could be combined with different binary classification tools to obtain accurate *de novo* plasmid reconstructions for multiple bacterial species. This means that gplas2 could, in theory, also be applied to the reconstruction of plasmids in metagenomic samples. However, since a greater number of plasmid-predicted unitigs is expected on metagenomes, the construction of plasmid walks will probably require parallelization in order to keep the computation time within practical limits.

Although our method constitutes a considerable improvement of the reconstruction of ARG-plasmids, some limitations should be noted. First, gplas2 does not include insertion sequences (and other repeated elements) into plasmid predictions. This facilitates the process of finding plasmid walks with homogeneous coverages and simplifies the resulting plasmidome network. However, insertion sequences play an important role in the structure and genomic plasticity of plasmids [43], and they are frequently involved in the mobility of ARGs [9, 44, 45]. Additionally, the localization of these MGEs can influence the expression levels of ARGs [46, 47], thereby impacting the resulting resistance phenotypes. Consequently, including insertion sequence (IS) elements would certainly improve the completeness and relevance of plasmid predictions. Some graph-based plasmid reconstruction methods, such as HyAsP [48], include repeated elements into predictions. This tool also constructs plasmid walks, and uses coverage information to predict IS copy numbers, thus allowing the same IS to be present in multiple replicons. In the gplas algorithm, considering repeated elements during the construction of the plasmid walks would lead to more entangled plasmidome networks and would complicate the subsequent partitioning step. As an alternative, we could envision adding labels to unitigs after the binning step, and then implementing a label propagation algorithm on the original assembly graph to determine to which bin the different IS elements belong. A similar approach is implemented by the tool GraphBin2 [49], which refines binning results of metagenomics samples. A second disadvantage of our method is the formation of chimaeras, which are bins composed of nodes from distinct replicons. As previously mentioned, accurate identification of plasmid-derived nodes reduces the number of chromosome–plasmid chimaeras. However, preventing the formation of plasmid–plasmid chimaeras is more challenging, especially for isolates carrying multiple large plasmids with similar copy numbers. Separating these chimaeras could be possible with the use of a plasmid-backbone reference database.

To conclude, in this work we have presented a new plasmidome prediction tool, named plasmidEC, and optimized gplas to accurately bin predicted plasmid sequences. Compared to existing binary classifiers, plasmidEC achieves increased recall<sub>(contig)</sub>, especially for contigs that derive from ARG plasmids. The integration of plasmidEC with gplas2 substantially improved the reconstruction of ARG plasmids in *E. coli*. Our method exceeded the binning capacity of the reference-based method MOB-suite, while retaining similar accuracy<sub>(bp)</sub> values. The presented approach constitutes the best alternative to accurately predict and reconstruct ARG plasmids *de novo* in the absence of long-read data.

#### Funding information

This work was partially supported by ZonMW (The Netherlands) [541 003 005 to A.C.S.], the Netherlands Centre of One Health (NCOH Complex systems and metagenomics) and by DiSSeMINATE (LSHM19138). This collaboration project is co-funded by the PPP Allowance made available by

Health-Holland, Top Sector Life Sciences and Health, to stimulate public-private partnerships. This work was partially supported by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions (grant No. 801133 to S.A.-A).

#### Author contributions

Conceptualization, J.A.P., A.C.S., S.A.A.; methodology, J.A.P., L.V., J.J.K., S.A.A.; validation and formal analysis, J.A.P., L.V., J.J.K.; resources, supervision and project administration, A.C.S., S.A.A., R.J.L.W., N.L.P.; data curation, J.A.P., L.V.; writing—original draft preparation, J.A.P.; writing—review and editing, J.A.P., A.C.S., N.L.P.; visualization, J.A.P., L.V. All authors have read and agreed to the published version of the manuscript.

#### Conflicts of interest

The authors declare that there are no conflicts of interest.

#### References

- Kern WV, Rieg S. Burden of bacterial bloodstream infection—a brief update on epidemiology and significance of multidrug-resistant pathogens. *Clin Microbiol Infect* 2020;26:151–157.
- Day MJ, Doumith M, Abernethy J, Hope R, Reynolds R, et al. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *J Antimicrob Chemother* 2016;71:2139–2142.
- Tumbarello M, Sanguinetti M, Montuori E, Trecarichi EM, Posteraro B, et al. Predictors of mortality in patients with bloodstream infections caused by extended-spectrum-beta-lactamase-producing *Enterobacteriaceae*: importance of inadequate initial antimicrobial treatment. *Antimicrob Agents Chemother* 2007;51:1987–1994.
- Mediavilla JR, Patrawalla A, Chen L, Chavda KD, Mathema B, et al. Colistin- and carbapenem-resistant *Escherichia coli* harboring mcr-1 and blaNDM-5, causing a complicated urinary tract infection in a patient from the United States. *mBio* 2016;7:e01191–16.
- Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022;399:629–655.
- Jiang X, Ellabaan MMH, Charusanti P, Munck C, Blin K, et al. Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat Commun* 2017;8:15784.
- Lerminiaux NA, Cameron ADS. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can J Microbiol* 2019;65:34–44.
- McInnes RS, McCallum GE, Lamberte LE, van Schaik W. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Curr Opin Microbiol* 2020;53:35–43.
- Che Y, Yang Y, Xu X, Brinda K, Polz MF, et al. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc Natl Acad Sci U S A* 2021;118:e2008731118.
- Zhang S, Abbas M, Rehman MU, Huang Y, Zhou R, et al. Dissemination of antibiotic resistance genes (ARGs) via integrons in *Escherichia coli*: a risk to human health. *Environ Pollut* 2020;266:115260.
- Norman A, Hansen LH, Sørensen SJ. Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc Lond B Biol Sci* 2009;364:2275–2289.
- Lopatkin AJ, Meredith HR, Srimani JK, Pfeiffer C, Durrett R, et al. Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat Commun* 2017;8:1689.
- von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, et al. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front Microbiol* 2016;7:173.
- Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *Elife* 2020;9:e53886.
- Bosch T, Lutgens SPM, Hermans MHA, Wever PC, Schneeberger PM, et al. Outbreak of NDM-1-producing *Klebsiella pneumoniae* in a Dutch hospital, with interspecies transfer of the resistance plasmid and unexpected occurrence in unrelated health care centers. *J Clin Microbiol* 2017;55:2380–2390.
- Acman M, van Dorp L, Santini JM, Balloux F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat Commun* 2020;11:1–11.
- Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* 2020;11:3602.
- 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:e000128.
- Paganini JA, Plantinga NL, Arredondo-Alonso S, Willems RJL, Schürch AC. Recovering *Escherichia coli* plasmids in the absence of long-read sequencing data. *Microorganisms* 2021;9:1613.
- Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4:e000206.
- Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, et al. gplasm: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics* 2020;36:3874–3876.
- Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;26:1721–1729.
- 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:e000211.
- van der Graaf-van Bloois L, Wagenaar JA, Zomer AL. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microb Genom* 2021;7:000683.
- 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. *Microb Genom* 2020;6:mgen000398.
- 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:e000224.
- 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.
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
- 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.
- Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 2020;75:3491–3500.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013. <http://arxiv.org/abs/1303.3997>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.

34. Arredondo-Alonso S, Top J, McNally A, Puranen S, Pesonen M, *et al.* Plasmids shaped the recent emergence of the major nosocomial pathogen *Enterococcus faecium*. *mBio* 2020;11:e03284-19.
35. Li Y, Luo Y. Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quant Biol* 2020;8:347–358.
36. Millán Arias P, Alipour F, Hill KA, Kari L, Chen C-H. DeLUCS: deep learning for unsupervised clustering of DNA sequences. *PLoS One* 2022;17:e0261531.
37. Wattanapornprom W, Thammarongtham C, Hongsthong A, Lertampaiporn S. Ensemble of Multiple Classifiers for Multilabel Classification of Plant Protein Subcellular Localization. *Life* 2021;11:293.
38. Xue T, Zhang S, Qiao H. i6mA-VC: a multi-classifier voting method for the computational identification of DNA N6-methyladenine sites. *Interdiscip Sci* 2021;13:413–425.
39. Douarre P-E, Mallet L, Radomski N, Felten A, Mistou M-Y. Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of IncF plasmids. *Front Microbiol* 2020;11:483.
40. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, *et al.* Niche and local geography shape the pangenome of wastewater- and livestock-associated *Enterobacteriaceae*. *Sci Adv* 2021;7:eabe3868.
41. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257.
42. Gomi R, Wyres KL, Holt KE. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb Genom* 2021;7:000550.
43. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit Rev Microbiol* 2017;43:709–730.
44. Razavi M, Kristiansson E, Flach C-F, Larsson DGJ. The association between insertion sequences and antibiotic resistance genes. *mSphere* 2020;5:e00418-20.
45. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev* 2018;31:e00088-17.
46. Kamruzzaman M, Patterson JD, Shoma S, Ginn AN, Partridge SR, *et al.* Relative strengths of promoters provided by common mobile genetic elements associated with resistance gene expression in gram-negative bacteria. *Antimicrob Agents Chemother* 2015;59:5088–5091.
47. Turton JF, Ward ME, Woodford N, Kaufmann ME, Pike R, *et al.* The role of ISAbal in expression of OXA carbapenemase genes in *Acinetobacter baumannii*. *FEMS Microbiol Lett* 2006;258:72–77.
48. Müller R, Chauve C. HyAsP, a greedy tool for plasmids identification. *Bioinformatics* 2019;35:4436–4439.
49. Mallawaarachchi VG, Wickramarachchi AS, Lin Y. Improving metagenomic binning results with overlapped bins using assembly graphs. *Algorithms Mol Biol* 2021;16:3.

**The Microbiology Society is a membership charity and not-for-profit publisher.**

**Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.**

**Find out more and submit your article at [microbiologyresearch.org](https://microbiologyresearch.org)**