On the (im)possibility to reconstruct plasmids from whole genome short-read sequencing data

Sergio Arredondo-Alonso[1], Willem van Schaik[1], Rob J. Willems[1], Anita C. Schürch[1*],

1 University Medical Center Utrecht - Department of Medical Microbiology

*** a.c.schurch@umcutrecht.nl**

Abstract

Plasmids are autonomous extra-chromosomal elements in bacterial cells that can carry genes that are important for bacterial survival. There is considerable interest in the automated reconstruction of plasmid sequences from short-read whole genome sequence (WGS) data. To benchmark algorithms for automated plasmid sequence reconstruction, we selected 42 publicly available complete bacterial genome sequences with associated sequencing reads from 12 genera, containing 148 plasmids. We predicted plasmids from WGS with four different programs (PlasmidSPAdes, Recycler, cBar and PlasmidFinder) and compared the outcome to the reference sequences. Recall and precision were calculated to measure the completeness and accuracy of each prediction.

PlasmidSPAdes reconstructs plasmids based on coverage differences in the assembly graph. It reconstructed most of the reference plasmids (recall = 0.82) with approximately a quarter of the predicted sequences corresponding to false positives (precision = 0.76). A total of 83.1 % of the reconstructions from genomes with multiple plasmids were merged and manual steps were necessary to separate individual plasmid sequences. Recycler searches the assembly graph for sub-graphs corresponding to circular sequences. It correctly predicted small plasmids but failed with long plasmids (recall = 0.12, precision = 0.28). cBar, which applies pentamer frequency composition analysis to detect plasmid-derived contigs, showed an overall recall and precision of 0.77 and 0.63. However, cBar only categorizes contigs as plasmid-derived and does not bin the different plasmids correctly within a bacterial isolate. PlasmidFinder, which searches for matches in a replicon database, had the highest precision (1.0) but was restricted by the contents of its database and the contig length obtained from *de novo* assembly (recall = 0.33). Based on this analysis we conclude that without long read information, plasmid reconstruction from WGS remains challenging and error-prone.

Introduction

Plasmids are a major driver of variation and adaptation in bacterial populations. The dissemination of multidrug resistance via transfer of plasmids leads to new antibiotic resistant bacteria such as *Escherichia coli* producing extended-spectrum beta-lactamases [1] or vancomoycin resistant *Enterococcus faecium* causing nosocomial outbreaks [2]. The prevalence of a plasmid in a bacterial population can increase due to environmental pressures include the presence of an antibiotic, but may cause a decrease in bacterial fitness in absence of selective pressure [3].

A bacterial cell can hold no, one or multiple plasmids with varying sizes and copy numbers. Traditionally, plasmid sequencing involved the extraction of plasmids using methods to specifically purify plasmid DNA, followed by shot-gun sequencing of the purified plasmid, which frequently necessitated closing of gaps by PCR or primer-walking [4]. Plasmid DNA purification is exceedingly difficult if it involves plasmids ranging from 50 kbp to 200 kbp [4, 5]. Alternatively, plasmid sequences can be assembled from whole genome sequencing data (WGS) sequenced by high throughput methods. However, plasmids often contain repeated sequences shared between the different physical DNA units of the genome, which prohibits complete assembly from short read data. Assembly often results in many fragmented contigs per genome of which their origin, plasmid or chromosome, is unclear [6]. Assembly alone is therefore insufficient to determine the origin of a contig and to differentiate contigs belonging to different plasmids. Recently, attempts to reconstruct plasmids from WGS data were automated in a number of programmes. Here, we benchmarked currently available programmes to detect and reconstruct plasmid sequences from short read sequencing data, starting either from the reads or from assembled contigs. The aim of this study was to determine whether it is possible to obtain complete plasmid sequences with state-of-the-art tools without manual expert intervention.

Programmes

Currently available plasmid reconstruction programmes either aim to determine whether a previously assembled contig is obtained from a plasmid (PlasmidFinder, cBAR), or try to reconstruct whole plasmid sequences from the (mapped) sequencing reads or the assembly graph (Recycler, PlasmidSPAdes, PLACNET) (Table 1).

One of the most widely used tools for plasmid detection and classification is a web tool called PlasmidFinder, developed to detect replicon sequences [7]. Two plasmids sharing the same replication mechanism cannot coexist in the long term within the same cell thus replicon sequences are used to classify plasmids into different incompatibility groups [8]. We downloaded the PlasmidFinder database containing 121 replicon sequences (updated on 16 March 2016) from the Center for Genomic Epidemiology (https//cge.cbs.dtu.dk//services/data.php). Contigs generated with SPAdes 3.8.2 [9] on a high performance computing cluster running CentOS7 were identfied as plasmids if they had a minimum identity of 80% and covered at least 60% of the replicon sequence [7]. For this purpose, we performed several nucleotide BLAST (NCBI-BLAST version 2.2.28+) searches against the PlasmidFinder database, to reproduce the results that would be obtained by the PlasmidFinder web-tool.

Unsupervised binning using differences in k-mer composition has been widely used in shotgun metagenomic algorithms [10–12]. Composition-based classification methods allow the clustering of contigs into distinct genomes and perform a species-level classification. Most of these methods are however not designed for application to isolated strains and do not report a classification between plasmid or chromosomal contigs. cBar was selected because it was specifically designed to predict plasmid-derived sequences based on differences in k-mer composition [13]. It relies on differences in pentamer frequencies from 881 complete prokaryotic sequences and gives a binary classification of chromosome- or plasmid-derived contigs. cBar version 1.2 was downloaded at http://csbl.bmb.uga.edu/ ffzhou/cBar/cBar.1.2.tar.gz and used to categorize contigs derived by SPAdes 3.8.2.

Plasmid constellation network (PLACNET) reconstructs plasmids from WGS by integrating three lines of evidence: (i) scaffold linking and coverage information from genome assembly, (ii) presence of replication initiator proteins (Rip) and relaxase proteins (Rel), (iii) similarity of the sequences with a custom database containing non-redundant plasmid sequences from NCBI [14]. PLACNET merges all the information into a single network where each component corresponds to a physical DNA unit. Repetitive sequences such as transposases or insertion sequences (IS) with a higher coverage are shared between components. Manual pruning in Cytoscape is necessary to duplicate and split the graph to obtain disjoint components in the final network [15]. Prediction reproducibility rates highly depend on the expertise of the researcher. As we aimed to test fully automated methods for plasmid reconstruction, we excluded PLACNET from the comparison.

More recently, two algorithms that reconstruct plasmids on basis of the information contained in the *de Bruijn* graph were developed: Recycler [16] and PlasmidSPAdes [17].

Recycler extracts the information from the de Bruijn graph searching for sub-graphs (cycles) corresponding to plasmids. Selection of the cycles is based on the following assumptions: (i) nodes forming a plasmid have a uniform coverage, (ii) a minimal path must be selected between edges because of repetitive sequences, (iii) contigs belonging to the same cycle have concordant read-end paired information and (iv) plasmid cycles exceed a minimum length [16]. For each sample, the assembly graph and resulting contigs corresponding to the maximum *k-mer* used by SPAdes 3.8.2 were selected. The BAM file required as input by Recycler was created by alignment of the trimmed reads against the resulting contigs using Bwa 0.7.12 [18] and samtools 1.3.1 [19].

PlasmidSPAdes assumes a highly uniform coverage of the contigs within the chromosome. It calculates the median coverage from the SPAdes assembly graph to estimate the chromosome coverage. By default, only contigs longer than 10 kbp are considered because repeated sequences are mostly present in shorter contigs and long contigs have a lower coverage variance. Contigs are classified as chromosomal edges if their coverage does not exceed a maximum deviation (default 0.3) from the median coverage. PlasmidSPAdes iteratively removes long chromosomal edges to transform the assembly graph into a plasmid graph. Finally, connected components in the plasmid graph are reported as putative plasmids [17].

Test data

To measure the performance of the different programmes on a range of bacterial species we selected 42 complete genome sequences from twelve different genera: *Aeromonas*, *Bacillus*, *Burkholderia*, *Citrobacter*, *Corynebacterium*, *Enterobacter*, *Escherichia*, *Klebsiella*, *Kluyvera*, *Providencia*, *Rhodobacter* and *Serratia* (Table 2). In total, the test data contained 148 plasmid sequences ranging from 1.55 kbp to 338.85 kbp and 45 chromosomal sequences from 0.93 Mbp to 6.26 Mbp (Figure 1).

All strains were previously sequenced by Pacific Biosystems PacBio RS II and Illumina Miseq or Hiseq paired-end libraries. Complete genome sequences were downloaded from GenBank and reads from the NCBI Sequence Read Archive (SRA) (Table 2). Low-quality bases at both ends of the reads were trimmed using the phred algorithm established by default in seqtk (version: 1.0-r31, github.com/lh3/seqtk.git).

Burkholderia cenocepacia DDS 22E-1 was included as a negative control. It contained three chromosomes with a length of 1.17 Mbp, 3.21 Mbp and 3.67 Mbp but no plasmid (Figure 1). The most complex composition of plasmids was present in *Klebsiella oxytoca* CAV1374 with a single chromosome and eleven plasmids ranging from 1.91 kbp to 332.95 kbp (Figure 1). In contrast, *Bacillus subtilis* subsp. natto BEST195 contained a single plasmid with a length of 5.84 kbp (Figure 1). This genome, along with *Corynebacterium callunae* DSM 20147 and *Enterococcus faecium* strain ATCC 700221, were the only gram-positive organisms included in the study.

Five genomes (*Escherichia coli* JJ1886 ; *Rhodobacter sphaeroides* 2.4.1, *Citrobacter freundii* CFNIH1, *Burkholderia cenocepacia* strain DDS 22E-1 and *C. callunae* DSM 20147) were previously used to validate Recycler and/or PlasmidSPAdes [16, 17]. These were selected to replicate the results described in the original publications (see Supplementary Data A.2).

Measures for the evaluation

We evaluated the performance of each programme regarding accuracy and completeness compared to i) coverage against each reference plasmid separately and ii) the whole reference genome. For Recycler, sequences considered were the cycles that were the output of the programme. For PlasmidSPAdes, we considered the connected components that were reported as putative plasmid sequences. For PlasmidFinder and cBAR, we considered the full length of the contigs that were predicted as either containing a replicon sequence (PlasmidFinder) or that was classified as plasmid based on its pentamer frequency (cBAR).

Quast 4.1 [20] was used to map the reconstructions against the reference chromosomes and plasmids using Nucmer alignments. We defined the following relevant values to evaluate the predictions:

- **Coverage** of each plasmid by the prediction. Defined as percentage of aligned bases of each prediction per genome against each reference plasmid, as reported as “Genome fraction” by Quast.

- **Reference plasmid frequency** as defined by the sum of the length of sequences which were true positive predictions (sequences mapping to reference plasmids) divided by the total length of the predicted sequences. Predicted sequences that mapped to both the reference plasmids and to the chromosome were considered as true positive results.

- **Chromosome frequency** as defined by the length of the sequences identified as false positive predictions, thus corresponding to the chromosome, divided by the total output sequence length. This can include non-plasmid mobile genetic elements such as phage or transposable elements.

- **Precision** was calculated to measure the accuracy of each prediction as the *reference plasmid frequency* divided by the sum of the *reference plasmid frequency* and *chromosome frequency.* Sequences not mapping to the reference genomes were excluded. Precision values of 1.0 indicated the absence of false positive predictions.

- **Recall** was calculated to measure the completeness of each prediction. The length of the sequences corresponding to true positive predictions was divided by the total reference plasmid length. This number was estimated using the genome fraction reported in Quast. Again, sequences not mapping to the reference genomes were excluded. A recall value of 1.0 indicated that all reference plasmids were predicted by the reconstruction. Lower recall values indicated the presence of false negative results.

- **Frequency of novel sequences** not mapping to the reference genomes. The sum of the length of reconstructed sequences not mapping to either the reference plasmids or the chromosome was divided by the total output length. These sequences were annotated using Prokka [21] and the annotation searched for genes corresponding to potential plasmid-located genes, such as Rip, Rel, Type IV components and toxin/antitoxin systems (TA). Furthermore, the sequences were compared to the non-redundant nucleotide database of the NCBI with BLAST. The best blast hit was extracted selecting minimum e-value and highest bit-score as previously described [16, 17]. A sequence match with a plasmid of a similar size suggested that the contig did not belong to a larger plasmid [22]. The completeness of the potential novel mobile elements was corroborated by generating a dot-plot mapping the sequence against itself using Gepard [23]. The presence of the same repeated sequence at the ends of the contig suggested a potential circularization signature.

- **Fraction of chromosome** wrongly predicted as plasmid sequences. This number was estimated using the genome fraction given by Quast selecting only the chromosome(s) of each genome.

Scaffold linkage of specific contigs in the PlasmidSPAdes assembly graph of a selection of genomes was visualized with Bandage [24]. Icarus [25] allowed the visualization of the alignments between the reference genomes and the predicted sequences.

The whole workflow was written in python2.7 and R (0.99.982-version) (available at git@gitlab.com:sirarredondo/Plasmid_Assembly.git) (Supplementary Figure 1).

Results

Reconstruction per plasmid

Out of 148 reference plasmids included in this study, 133 (89.9 %) were reconstructed by either PlasmidFinder, cBar, Recycler or PlasmidSPAdes with a coverage of each plasmid by the predicted plasmid sequences of at least 90 % (Figure 2). PlasmidSPAdes recovered 125 plasmids, cBar 84 plasmids, Recycler 21 plasmid and PlasmidFinder 13 plasmids at a coverage of 90 % or more. While the coverage ratio of reference plasmids by the predictions declined with plasmid size for Recycler, cBAR and PlasmidFinder predictions, it remained the same for PlasmidSPAdes predictions. Both programmes with a high average coverage of each plasmid by the prediction (PlasmidSPAdes and cBAR, 87.2 % and 85.5 %, respectively) did not, or incompletely, report plasmid boundaries. cBar predicted contigs as either "plasmid" or "chromosome" but did not sort the sequences into different plasmids (binning). PlasmidSPAdes merged plasmids in 83 % of all the genomes with several reference plasmids, and plasmid boundaries were not readily retrievable. For example, *Citrobacter freundii* CAV1321 had nine reference plasmids ranging from 1.9 kbp to 243.7 kbp (Figure 1). PlasmidSPAdes reconstructed a single component from the plasmid graph with a length of 479.1 kbp, which was composed of 27 contigs (>1 kbp) from the nine reference plasmids. Despite the lack of plasmid boundaries, the completeness of the prediction was outstanding with a recall value of 0.97. Therefore we further evaluated the performance of each programme on the genome level rather than on an individual plasmid level.

Reconstruction per genome

PlasmidSPAdes

A total of 18.2 Mbp was detected as plasmid sequences by PlasmidSPAdes with an average reference plasmid frequency of 0.72 and an average chromosome frequency of 0.22 (shown in Figure 3). Surprisingly, a frequency of 0.06 corresponding to sequences not mapping to the reference genomes was detected. We obtained an overall precision of 0.76 from PlasmidSPAdes while the overall recall of PlasmidSPAdes (0.82) indicated that the majority of plasmids were present in the prediction (Figure 3).

The overall chromosome recovery of PlasmidSPAdes was 0.07, indicating that erroneous assignment of chromosomal contigs to plasmids was not common. Despite this low value, if a chromosome contig was not removed from the initial assembly graph the frequency of false positive results (chromosome frequency) significantly increased. This situation was reflected in *Klebsiella pneumoniae* CAV1596 where PlasmidSPAdes predicted a component of 379.17 kbp as putative plasmid sequences. From this total value, 172.64 kbp were part of the chromosome representing a chromosome frequency of 0.46. The reported chromosome frequency often included mobile genetic elements such as transposases or prophages which were not removed from the assembly graph.

In some genomes, the recall obtained was lower than 0.20 such as *Klebsiella pneumoniae* KPN223 or *Corynebacterium callunae* DSM 20147. In addition, *Enterobacter faecium* ATCC 700221 showed the highest chromosome recovery with a value of 0.38. Further analysis of *E.faecium* ATCC 700221 suggested a non-uniform coverage along the chromosome, and, consequently, most of the contigs erroneously predicted were near the chromosomal origin of replication.

Two strains (*E. coli* JJ1886 and *E. coli* JJ1887) were further analyzed because they showed a high number of contigs not mapping to the reference genomes as shown by a frequency of novel plasmids of 0.38 and 0.91 respectively. The results suggested a contamination from *Staphylococcus aureus*, probably during the library preparation of *E. coli* JJ1886 and *E. coli* JJ1887. Both strains were part of the same NCBI BioProject (Table 2). The chromosome and plasmids of *S. aureus* were not removed from the graph given by SPAdes because their coverage differed from the *E. coli* chromosome coverage. This suggests that contaminants may interfere with plasmid reconstruction by PlasmidSPAdes.

Most of the novel sequences not mapping to the reference genomes were detected as isolated components by PlasmidSPAdes with an intermediate copy number as inferred from their coverage ratio. Components formed by a single contig and with a best blast hit corresponding to a plasmid or containing a plasmid-related gene were mapped against themselves by a dot-plot to infer circularity. To get the correct sequence from these putative novel plasmids, it was necessary to remove one of the repeated sequences present at the ends of the contig (Supplementary Data A.3).

Recycler

The total number of plasmid sequences predicted by Recycler was 3.07 Mbp (Figure 4). From the total predictions by Recycler we obtained a plasmid frequency of 0.24, a chromosome frequency of 0.62 and a frequency of contigs not mapping to the reference genomes of 0.14. This resulted in an overall precision of 0.28 indicating a high number of sequences originating from the chromosome.

Recycler obtained an overall recall of 0.12 and a chromosome recovery of 0.01. However, in strains with relatively small plasmids (*B. subtilis* subsp. natto BEST 195, *Enterobacter aerogenes* CAV1320 and

Providencia stuartii ATCC 33672 (Figure 4) with plasmids of 5.8 kbp, 13.9 kbp and 48.86 kbp) the recall value was 1.0. These plasmids were covered by single and circular contigs. Recycler is specifically designed to extract circular sequences from the assembly graph. In *Citrobacter freundii* CAV1741 and *Klebsiella oxytoca* CAV1099, Recycler detected several circular sequences, including two large reference plasmids of 100.8 kbp and 111.3 kbp.

Due to the circular nature of other mobile elements, such as phage genomes, Recycler was able to extract those as well. This was reflected in the genome projects *Enterobacter cloacae* strain CAV1311, *E.cloacae* strain CAV1411, *E.cloacae* strain CAV1668 and *E.cloacae* strain CAV1669. In these strains, Recycler obtained a precision of 0.0 because no reference plasmid sequences were extracted by the algorithm. However, Recycler extracted a phage sequence (41.9 kbp).

Most of the novel sequences which do not map to the reference genomes reconstructed by Recycler were also detected as isolated components by PlasmidSPAdes (Figure 4). Common features of these novel sequences are a length less than 10 kbp and an intermediate copy number (Supplementary Data A.3).

cBar

cBar predicted every contig as either plasmid-derived or chromosome-derived. In order to maintain comparability, we only considered sequences predicted as plasmid to measure the performance in each genome.

This resulted in an overall precision and recall of 0.63 and 0.77 respectively. A substantial amount of contigs corresponding to reference plasmids was recovered. For instance, *C. freundii* CAV1321 was previously highlighted because of its complexity (Figure 1) and low recall values obtained by PlasmidSPAdes and Recycler (Figures 3 and 4). cBar however obtained a recall value of 0.93 for this strain indicating a high completeness of the results. However, the precision varied largely across genomes, as reflected in *Providencia stuartii* ATCC 33762 which contains a single reference plasmid of 48.87 kbp. This plasmid was correctly detected by cBar obtaining a recall value of 1.0. Nevertheless, it wrongly predicted 19 contigs (>500 bp) as plasmids which mapped to the chromosome, resulting in a precision of 0.34 (Figure 5).

As shown in Figure 5, precision and recall value were 0.0 in *B. subtilis* subsp. natto BEST195 and *E. aerogenes* CAV1320. Those bacterial strains carry single plasmids that were assembled into a single contig (Figure 1). The algorithm, however, erroneously predicted those contigs as chromosome-derived.

PlasmidFinder

PlasmidFinder was able to detect at least one plasmid replicon sequence in 38 of the bacterial strains, but failed to detect any replicon sequence in *B. cenocepacia* DDS 22E-1, *P. stuartii* ATCC 33672, *R. sphaeroides* 2-4-1, *E. faecium* ATCC 700221 and *C. callunae* DSM 20147.

The overall precision of PlasmidFinder was 1.0, indicating that no false positive sequences were predicted as plasmids. However, the overall recall of 0.33 was due to the low completeness of the results as shown in Figure 6. The recall of PlasmidFinder was directly linked to the size of the contigs where the replicon sequence was detected. For instance, in *E. aerogenes* CAV1320 we obtained a recall value of 1.0 because the strain carried a single 14 kpb plasmid that was completely assembled into a single contig containing a replicon sequence.

Conclusions

We compared four different programmes to reconstruct or predict plasmid sequences from WGS data. The large majority of the sequences of the plasmids (89.9 %) could be reconstructed by one of the programmes when compared to the reference plasmids. However, in many cases, the reconstructions were fragmented (all programmes), contaminated by chromosome sequences (cBAR, Recycler, PlasmidSPAdes), boundaries of the plasmids were unclear (cBAR, PlasmidSPAdes) and plasmids incomplete (all programmes). In absence of reference plasmid sequences, disentangling or binning the reconstructions into separate plasmids is a challenging step that still has to be solved.

PlasmidSPAdes recovered 82.4 %, of the reference plasmids present in each genome. However, in many cases (83 % of all genome projects with more than one plasmid), several plasmids were merged into a single component, along with chromosomal sequences (on average 24 %). By visualizing the plasmid graph and connecting contigs with a similar coverage and scaffolding linkage, plasmid sub-graphs can, theoretically, be separated manually, if the different plasmids sufficiently differ in their copy number [17] (Supplementary Data A.2.4). A similar manual step was previously used in PLACNET [14] where manual pruning is necessary to duplicate repeated sequences such as transposases to split plasmids into different physical DNA units. However, whether manual interventions are successful is highly dependent on the expertise of the individual analyzing the data, can be difficult to reproduce independently and limits the high-throughput analysis of WGS data.

Recycler applies an innovative approach to plasmid reconstruction and succesfully extracted complete plasmid sequences if they had circular features. Most large plasmids however tend to be assembled into several contigs due to the presence of repeated sequences with high coverage. Recycler failed to extract these types of plasmids and in many cases only extracted mobile elements belonging to the chromosome. However, Recycler was also designed to detect plasmids in metagenomes, and may be useful to extract circular sequences from samples with variances in coverage.

To our surprise, PlasmidSPAdes and Recycler reconstructed 36 DNA fragments (>1 kbp) not present in the completed reference sequences. They had a length of less than 10 kbp and were composed by a single contig. These sequences could originate from sequences neglected or avoided in the reference assembly because they constituted contamination, but could also represent small DNA fragments not captured by the long read sequencing techniques, such as small cryptic plasmids. Small cryptic plasmids are mostly composed of genes involved in plasmid replication and were previously described in ESBL-producing *E.coli* [22]. A total of 19 putative small cryptic plasmids were extracted by Recycler. Consequently, Recycler may be a valuable tool to obtain whole sequences of short length plasmids from cultivated and uncultivated bacteria.

cBar was originally designed to categorize chromosome and plasmids in metagenomic sequences by comparing pentamer frequencies of a plasmid database. The accuracy of this approach is known to be lower for long plasmids because of similarities in nucleotide composition to the host chromosome [26]. However, the overall recall of cBar is high (0.78) and it might be well-suited to confirm if a sequence is plasmid-derived.

The results of PlasmidFinder showed an outstanding 1.0 true positive rate indicating a high reliability of the prediction. Being initially designed for Enterobacteriaceae, it was not able to detect any plasmid replication initiator protein in four bacterial strains including three gram-positive genomes. If applied to PlasmidSPAdes predictions, the detection of different incompatibility groups by PlasmidFinder could indicate the presence of two or more plasmids merged together into a single component.

In this study, plasmid reference sequences were present for comparison, something which is lacking in WGS projects for which these tools have been developed. The presence of repeated sequences shared in different physical DNA units, indiscriminate pentamer frequencies and similar coverage ratios make the *de novo* reconstruction of plasmids from WGS challenging, even with the help of the reconstruction programmes tested here. To obtain the full sequences of plasmids, long read sequencing data can be a solution [5]. Nonetheless, the comparably high costs of long read sequencing by Pacific Biosystems PacBio RS II or Oxford Nanopore Technologies Ltd and the relatively high error rate of these techniques make the combination with short-read sequencing data desirable. Moreover, *de novo* assembly using exclusively short-read sequencing data can identify contigs, potentially representing small plasmids, which are not covered by reads generated by long-read sequencing data. This may be crucial to identify the entirety of the plasmids repertoire and, with that, obtain complete genome sequences.

References

[1] SS Mo, JS Slettemeås, ES Berg, M Norström, and M Sunde. Plasmid and Host Strain Characteristics of Escherichia coli Resistant to Extended-Spectrum Cephalosporins in the Norwegian Broiler Production. *PLoS ONE*, 11(4):e0154019, 2016.

[2] Ana R. Freitas, Ana P. Tedim, Maria V. Francia, Lars B. Jensen, Carla Novais, Luísa Peixe, Antonio Sánchez-Valenzuela, Arnfinn Sundsfjord, Kristin Hegstad, Guido Werner, Ewa Sadowy, Anette M. Hammerum, Lourdes Garcia-Migura, Rob J. Willems, Fernando Baquero, and Teresa M. Coque. Multilevel population genetic analysis of vanA and vanB Enterococcus faecium causing nosocomial outbreaks in 27 countries (1986–2012). *Journal of Antimicrobial Chemotherapy*, 2016.

[3] Alvaro San Milian, Alfonso Santos-Lopez, Rafael Ortega-Huedo, Cristina Bernabe-Balas, Sean P Kennedy, and Bruno Gonzalez-Zorn. Small-Plasmid-Mediated Antibiotic Resistance Is Enhanced by Increases in Plasmid Copy Number and Bacterial Fitness. *Antimicrobial Agents and Chemotherapy*, 59(6):3335–3341, 2015.

[4] Kornelia Smalla, Sven Jechalke, and Eva M Top. Plasmid detection, characterization and ecology. *Microbiology Spectr.*, 121(8):1265–1272, 2015.

[5] Sean Conlan, Pamela J Thomas, Clayton Deming, Morgan Park, Anna F Lau, John P Dekker, Evan S Snitkin, Tyson A Clark, Khai Luong, Yi Song, Yu-chih Tsai, Matthew Boitano, Jyoti Dayal, Shelise Y Brooks, Brian Schmidt, Alice C Young, James W Thomas, Gerard G Bouffard, Robert W Blakesley, NISC Comparative Sequencing Program, James C Mullikin, Jonas Korlach, David K Henderson, Karen M Frank, Tara N Palmore, and Julia A Segre. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Science translational medicine*, 6(254):254ra126, 2014.

[6] María De Toro, M Pilar Garcillán-Barcia, Fernando De, and La Cruz. Plasmid Diversity and Adaptation Analyzed by Massive Sequencing of Escherichia coli Plasmids. *Microbiol Spectrum*, 2(6):PLAS–0031, 2014.

[7] Alessandra Carattoli, Ea Zankari, Aurora Garciá-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Mløler Aarestrup, and Henrik Hasman. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, 58(7):3895–3903, 2014.

[8] María Pilar Garcillán-Barcia, María Victoria Francia, and Fernando De La Cruz. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiology Reviews*, 33(3):657–687, 2009.

[9] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey a. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max a. Alekseyev, and Pavel a. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[10] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, 2014.

[11] Zhong Wang, Dongwan D Kang, Jeff Froula, and Rob Egan. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, e1165, 2015.

[12] David R Kelley and Steven L Salzberg. Clustering metagenomic sequences with interpolated Markov models. *BMC bioinformatics*, 11(1):544, 2010.

[13] Fengfeng Zhou and Ying Xu. cBar: A computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16):2051–2052, 2010.

[14] Val F. Lanza, Maria de Toro, M. Pilar Garcillan-Barcia, Azucena Mora, Jorge Blanco, Teresa M. Coque, and Fernando de la Cruz. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. *PLoS Genetics*, 10(12):e1004766, 2014.

[15] Rowan Christmas, Iliana Avila-Campillo, Hamid Bolouri, Benno Schwikowski, Mark Anderson, Ryan Kelley, Nerius Landys, Chris Workman, Trey Ideker, Ethan Cerami, Rob Sheridan, Gary D. Bader, and Chris Sander. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

[16] Roye Rozov, Aya Brown Kav, David Bogumil, Eran Halperin, Itzhak Mizrahi, and Ron Shamir. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *BioRxiv*, page 029926, 2015.

[17] Dmitry Antipov, Nolan Hartwick, Max Shen, Mikhail Raiko, and Pavel A Pevzner. plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. *Bioinformatics*, 2016.

[18] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[19] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[20] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[21] Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[22] Alma Brolund, Oscar Franzén, Öjar Melefors, Karin Tegmark-Wisell, and Linus Sandegren. Plasmidome-Analysis of ESBL-Producing Escherichia coli Using Conventional Typing and High-Throughput Sequencing. *PLoS ONE*, 8(6):e65793, 2013.

[23] Jan Krumsiek, Roland Arnold, and Thomas Rattei. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028, 2007.

[24] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 2015.

[25] Alla Mikheenko, Gleb Valin, Andrey Prjibelski, Vladislav Saveliev, and Alexey Gurevich. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics*, 2016.

[26] Peter W. Harrison, Ryan P J Lower, Nayoung K D Kim, and J. Peter W Young. Introducing the bacterial 'chromid': Not a chromosome, not a plasmid. *Trends in Microbiology*, 18(4):141–148, 2010.

[27] Alvaro San Millan, Karl Heilbron, and R Craig MacLean. Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. *The ISME journal*, 8(3):601–12, 2014.

[28] Mayumi Kamada, Sumitaka Hase, Kengo Sato, Atsushi Toyoda, Asao Fujiyama, and Yasubumi Sakakibara. Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads. *PLoS ONE*, 9(10):e109999, 2014.

Tables and Figures

Table 1. Overview of programmes to reconstruct or predict plasmids from short read sequencing data.

	Input	**Paired-end information**	**Coverage**	**k-mer composition**	**de Bruijn graph**	**Similarity to replicons**	**Similarity to relaxases**	**Similarity to plasmids**	**Web-tool**	**Command-line interface**	**Included in study**
PlasmidFinder [7]	Contigs					✓			✓		✓
cBAR [13]	Contigs			✓						✓	✓
Recycler [16]	BAM+assembly graph	✓	✓		✓					✓	✓
PlasmidSPAdes [17]	Reads	✓	✓		✓					✓	✓
PLACNET [14]	BAM/SAM+contigs	✓	✓			✓	✓	✓		✓	

Table 2. SRA and Bioproject accessions of each genome used in this study.

Strain	**SRA**	**Bioproject**
Aeromonas veronii strain AVNIH1	SRR3465535	PRJNA279607
Bacillus subtilis subsp. natto BEST195	DRR016448	PRJDA38027
Burkholderia cenocepacia strain DDS 22E-1	SRR1618480	PRJNA244014
Citrobacter freundii CFNIH1	SRR1284629	PRJNA202883
Citrobacter freundii strain CAV1321	SRR2965690	PRJNA246471
Citrobacter freundii strain CAV1741	SRR2965739	PRJNA246471
Corynebacterium callunae DSM 20147	SRR892039	PRJNA185570
Enterobacter aerogenes strain CAV1320	SRR2965748	PRJNA246471
Enterobacter asburiae strain CAV1043	SRR2965752	PRJNA246471
Enterobacter cloacae ECNIH2	SRR1515967	PRJNA202893
Enterobacter cloacae ECNIH3	SRR1576778	PRJNA202894
Enterobacter cloacae ECR091	SRR1576808	PRJNA202892
Enterobacter cloacae strain CAV1311	SRR2965815	PRJNA246471
Enterobacter cloacae strain CAV1411	SRR2965820	PRJNA246471
Enterobacter cloacae strain CAV1668	SRR2965612	PRJNA246471
Enterobacter cloacae strain CAV1669	SRR2965616	PRJNA246471
Enterococcus faecium strain ATCC 700221	SRR3176159	PRJNA311738
Escherichia coli JJ1886	SRR933487	PRJNA211153
Escherichia coli JJ1887	SRR933489	PRJNA211153
Escherichia coli strain Eco889	SRR3465539	PRJNA279654
Klebsiella oxytoca KONIH1	SRR1501122	PRJNA202895
Klebsiella oxytoca strain CAV1099	SRR2965639	PRJNA246471
Klebsiella oxytoca strain CAV1335	SRR2965660	PRJNA246471
Klebsiella oxytoca strain CAV1374	SRR2965655	PRJNA246471
Klebsiella pneumoniae strain AATZP	SRR3228444	PRJNA279650
Klebsiella pneumoniae strain CAV1193	SRR2965672	PRJNA246471
Klebsiella pneumoniae strain CAV1344	SRR1582875	PRJNA246471
Klebsiella pneumoniae strain CAV1392	SRR1582895	PRJNA246471
Klebsiella pneumoniae strain CAV1596	SRR1582868	PRJNA246471
Klebsiella pneumoniae strain Kpn223	SRR3465557	PRJNA279655
Klebsiella pneumoniae strain Kpn555	SRR3465562	PRJNA279656
Klebsiella pneumoniae strain KPNIH36	SRR3222156	PRJNA284365
Klebsiella pneumoniae strain KPNIH39	SRR3217430	PRJNA279611
Klebsiella pneumoniae strain PMK1	SRR1508819	PRJNA253300
Klebsiella pneumoniae subsp. pneumoniae KPNIH1	SRR1505904	PRJNA73191
Klebsiella pneumoniae subsp. pneumoniae KPNIH10	SRR1427234	PRJNA73843
Klebsiella pneumoniae subsp. pneumoniae KPNIH24	SRR1501128	PRJNA173233
Klebsiella pneumoniae subsp. pneumoniae KPNIH27	SRR1427243	PRJNA198783
Kluyvera intermedia strain CAV1151	SRR2965721	PRJNA246471
Providencia stuartii strain ATCC 33672	SRR1558174	PRJNA244575
Rhodobacter sphaeroides 2.4.1	SRR522246	PRJNA40077
Serratia marcescens strain CAV1492	SRR2965730	PRJNA246471

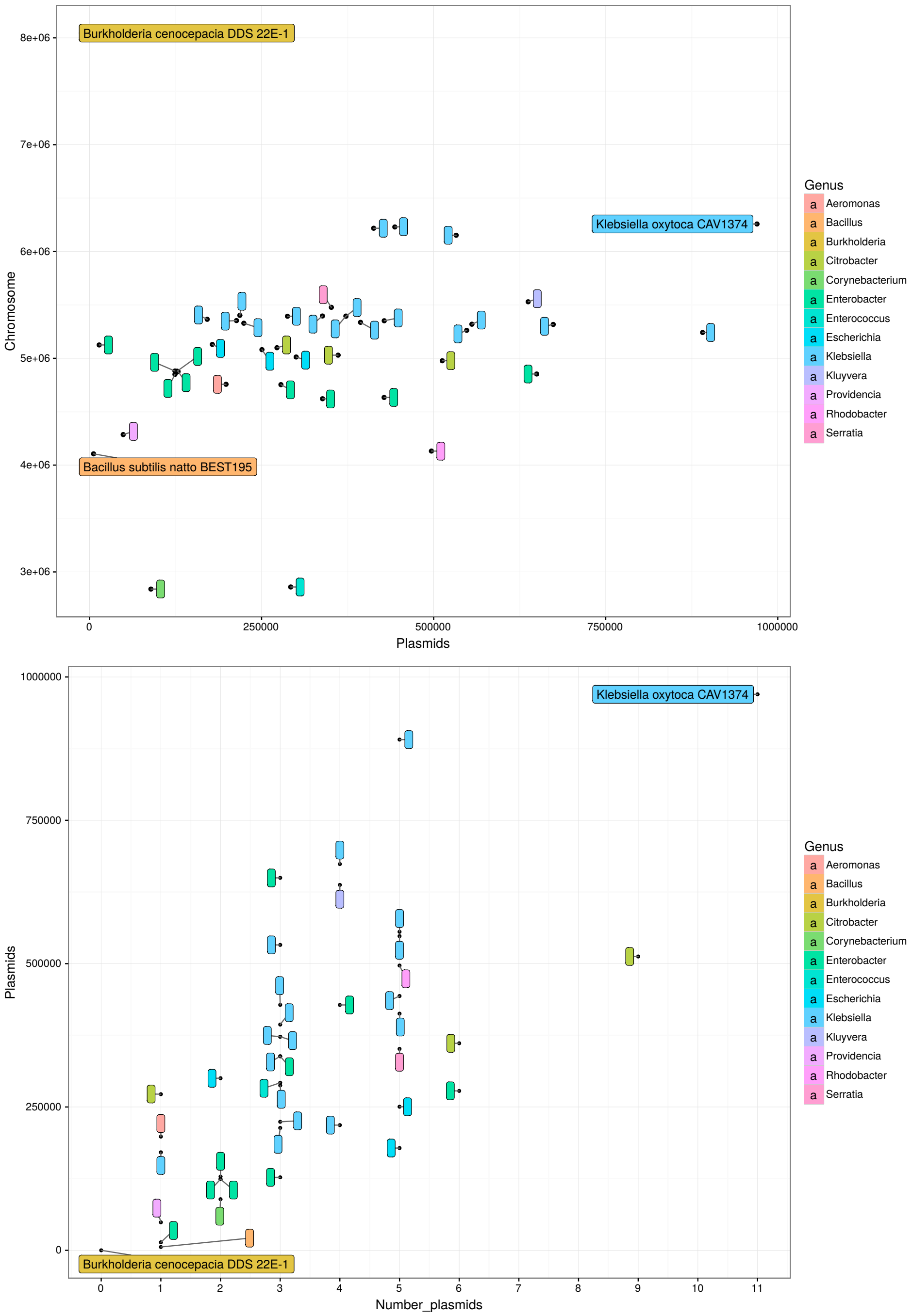

Figure 1. Scatter plots of (top) the total length of reference plasmids versus the chromosome length of each bacterial genome and (bottom) the total number of reference plasmids versus the total plasmid length per genome. Different genera are represented with colored boxes attached to data points with arrows. Species described in the text are highlighted and their full name given.

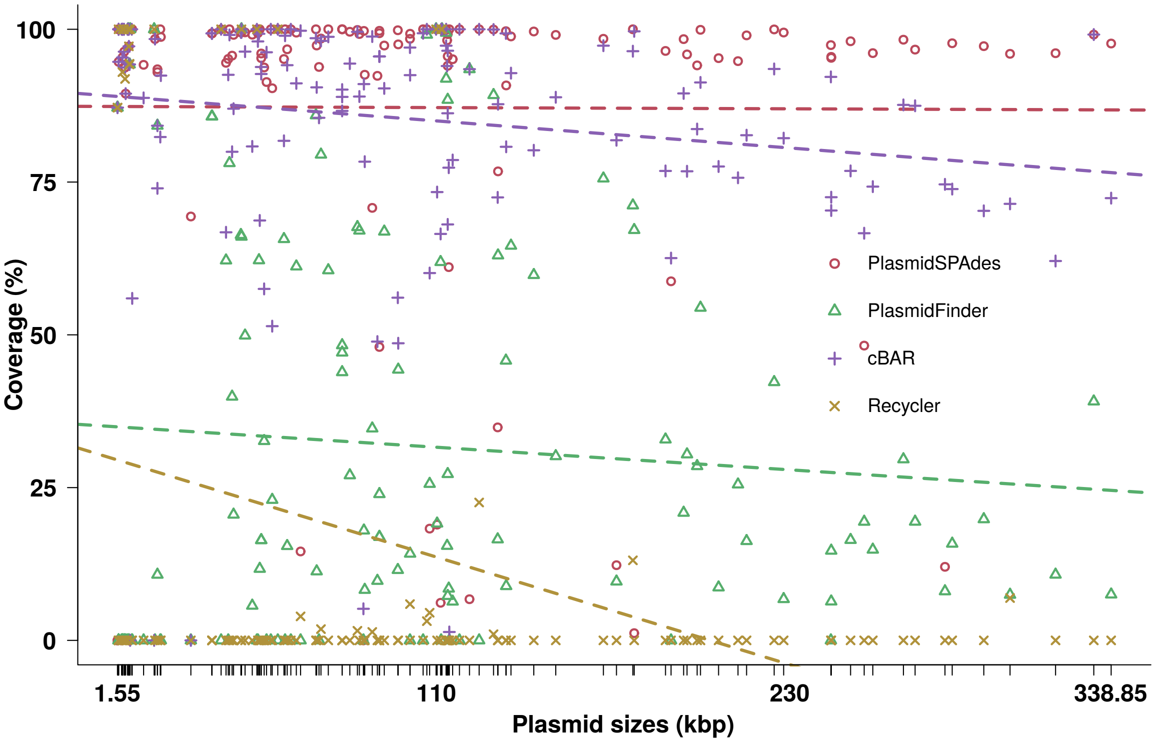

Figure 2. Coverage of reference plasmids by predicted plasmid sequences from PlasmidSPAdes, PlasmidFinder, cBAR and Recycler. Coverage was calculated by aligning the reference plasmid sequences against the plasmid predictions of each genome and disregarded plasmid binning (if any). Lines indicate linear least squares regression fits to data points. Tick marks on the x-axis represent plasmid sizes.

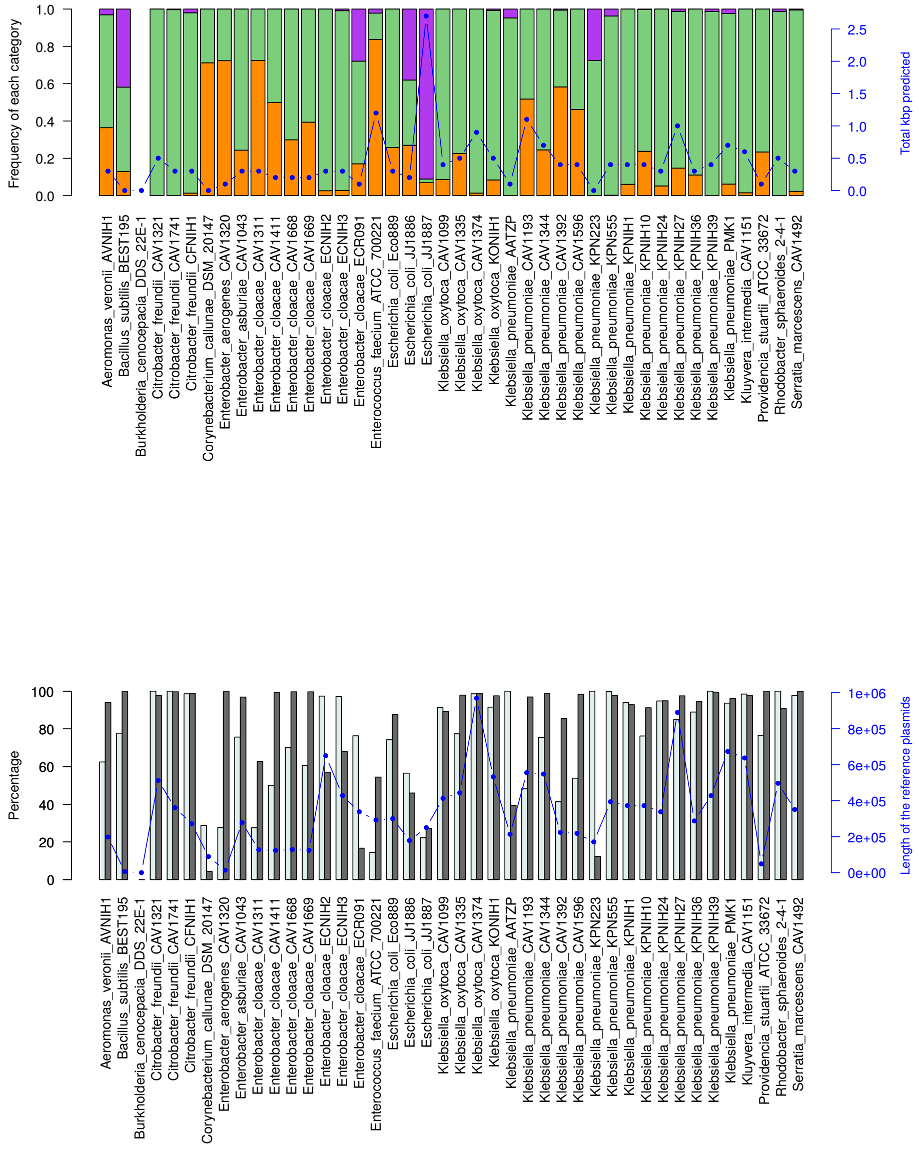

Figure 3. Performance of PlasmidSPAdes per genome. Top: As plasmids predicted sequences that map to reference plasmids (green), to the reference chromosome (orange) or to neither the reference chromosome or the reference plasmids (violet). On the right y-axis the total length (in kbp) of reconstructed plasmid sequences is indicated. Bottom: Precision (white) and recall (gray) values per genome. The total reference plasmid length is indicated on the y-axis.

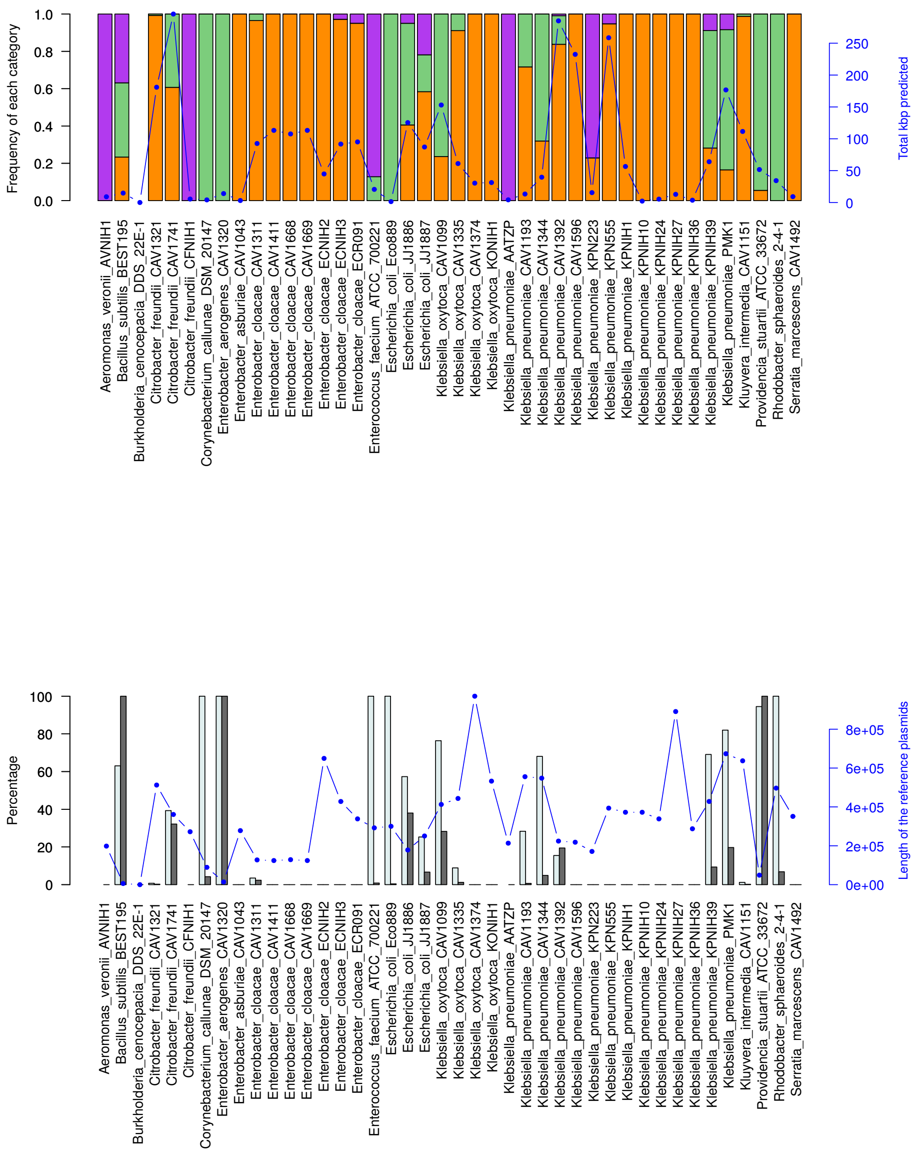

Figure 4. Performance of Recycler per genome. Top: As plasmids predicted sequences that map to reference plasmids (green), to the reference chromosome (orange) or to neither the reference chromosome or the reference plasmids (violet). On the right y-axis the total length (in kbp) of reconstructed plasmid sequences is indicated. Bottom: Precision (white) and recall (gray) values per genome. The total reference plasmid length is indicated on the y-axis.

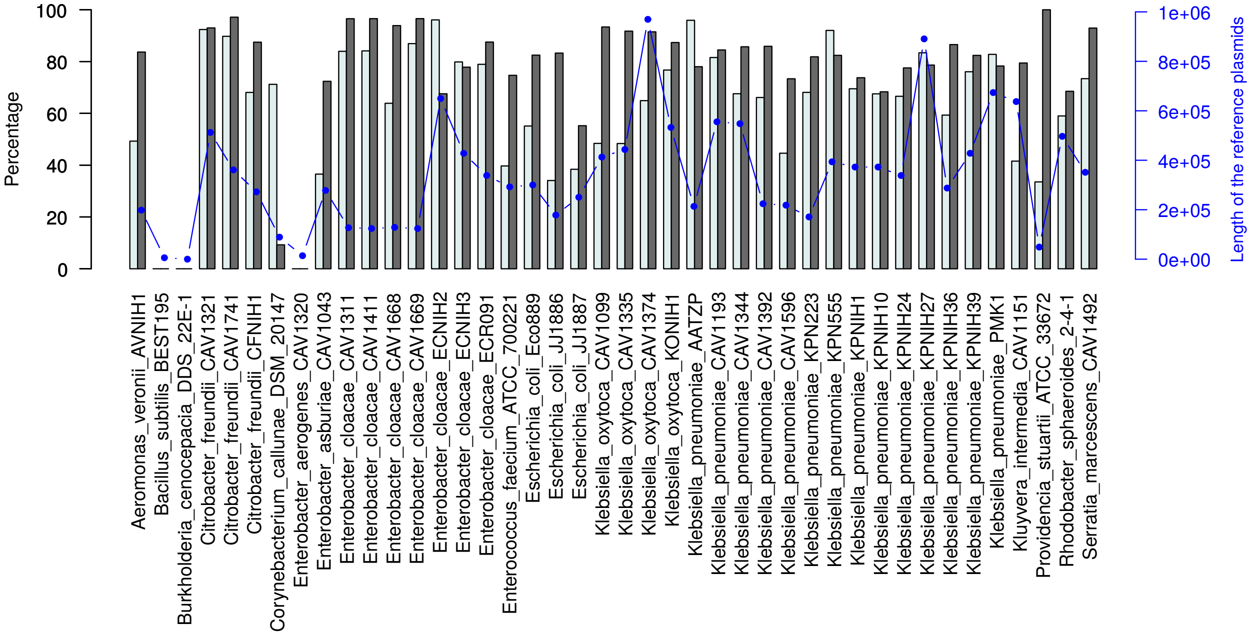

Figure 5. Performance of cBar per genome. Precision and recall values are represented in white and gray bars respectively. Precision and recall values of 100 (in percentage) indicate maximum completeness and exactness. The total reference plasmid length is indicated on the y-axis.

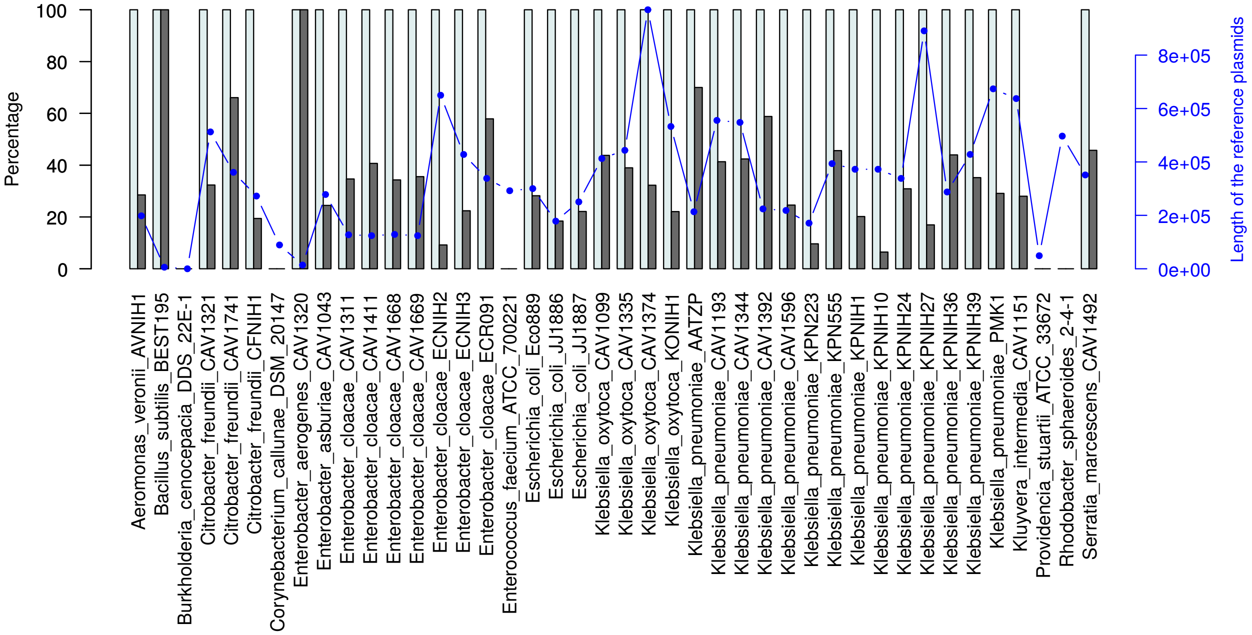

Figure 6. Performance of PlasmidFinder per genome. Precision and recall values are represented in white and gray bars respectively. Precision and recall values of 100 (in percentage) indicate maximum completeness and exactness. The total reference plasmid length is indicated on the y-axis.

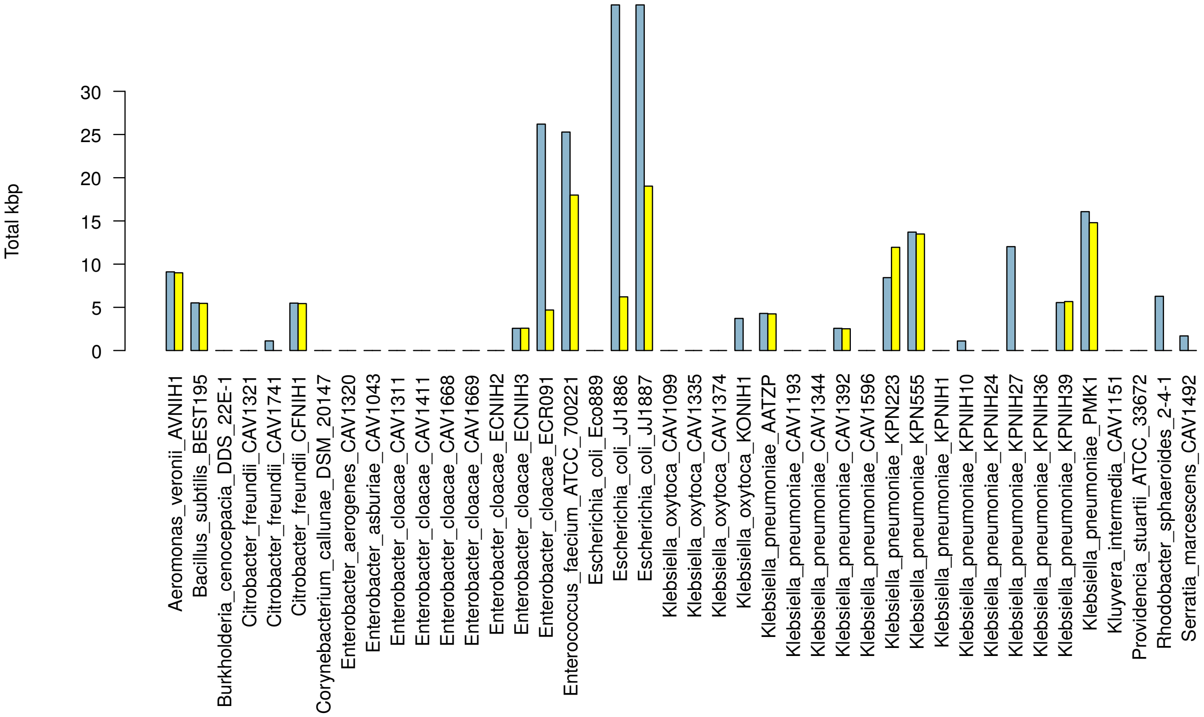

Figure 7. Predicted sequences not mapping to reference plasmids or chromosome as predicted by PlasmidSPAdes (blue) and Recycler (yellow). PlasmidSPAdes detected 2.44 Mbp and 82.32 kbp corresponding to *E.coli* JJ1887 and *E.coli* JJ1886 respectively.