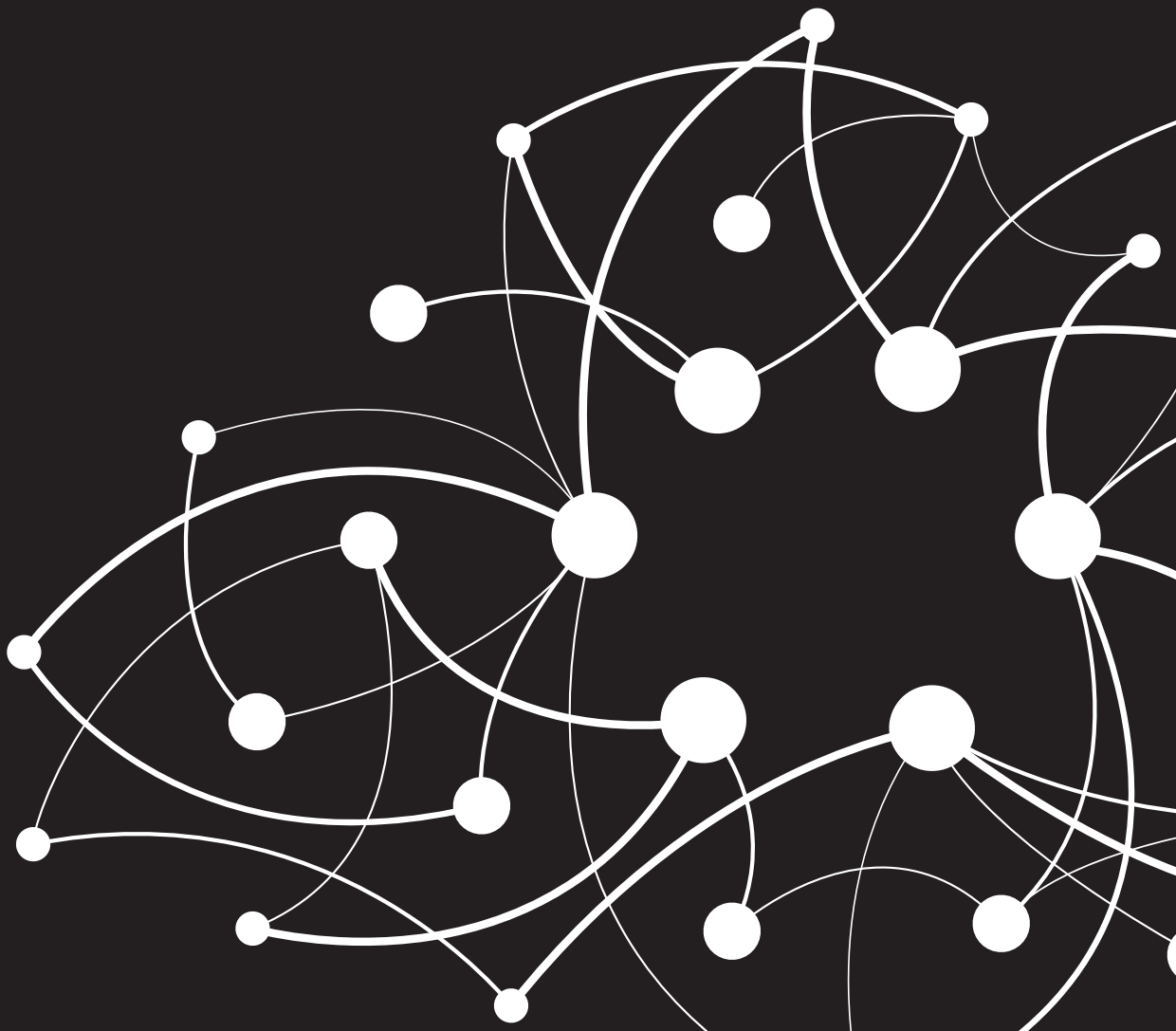


# Reconstructing AMR plasmids

New methods to track the dissemination of mobile genetic elements in ICUs

Julián Andrés Paganini





# **Reconstructing AMR plasmids**

New methods to track the dissemination  
of mobile genetic elements in ICUs

Julián Andrés Paganini

Reconstructing AMR Plasmids: New methods to track the dissemination of mobile genetic elements in ICUs. PhD thesis, Utrecht University, the Netherlands

Author: Julián Andrés Paganini

Cover design: Mara Mondini

Lay-out: Mara Mondini

Printed by: ProefschriftMaken | [www.proefschriftmaken.nl](http://www.proefschriftmaken.nl)

ISBN: 978-94-6469-450-5

© Julián Andrés Paganini Utrecht, the Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in an online retrieval system or transmitted in any form or by any means without permission of the author. The copyright of the articles that have been published has been transferred to the respective journals.

Printing of this thesis was financially supported by the University Medical Center Utrecht, the Netherlands Society of Medical Microbiology (NVMM) and the Royal Netherlands Society for Microbiology (KNVM)

# Reconstructing AMR plasmids

New methods to track the dissemination  
of mobile genetic elements in ICUs

**AMR-plasmiden reconstrueren**  
Nieuwe methoden om de verspreiding van  
mobiele genetische elementen op IC's te volgen

(met een samenvatting in het Nederlands)

## Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag  
van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit  
van het college voor promoties in het openbaar te verdedigen

op woensdag 27 september 2023 des middags te 12:15 uur

door

**Julián Andrés Paganini**

geboren op 7 september 1988  
te Rosario, Santa Fe, Argentinië

**Promotor:**

Prof. Dr. R.J.L. Willems

**Copromotoren:**

Dr. A.C. Schürch

Dr. N.L. Plantinga

**Beoordelingscommissie:**

Prof. Dr. M.J.M. Bonten (voorzitter)

Prof. Dr. B.E. Dutilh

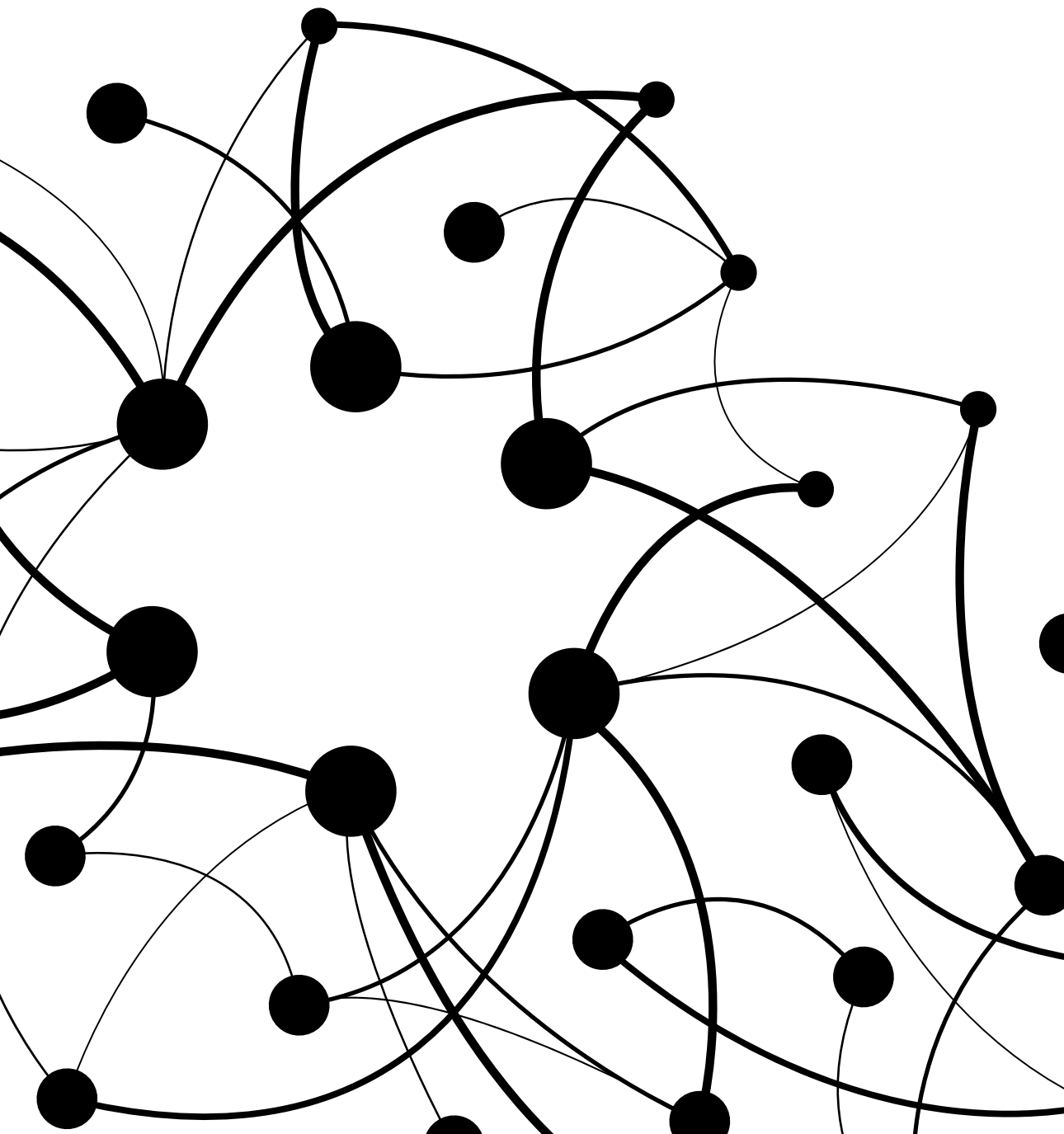
Prof. Dr. F. Hagen

Prof. Dr. C. Schultsz

Prof. Dr. B. Snel

## Table of Contents

07	<b>Chapter 1</b> General Introduction
25	<b>Chapter 2</b> Recovering <i>Escherichia coli</i> plasmids in the absence of long-read data
63	<b>Chapter 3</b> An optimised short-read approach to predict and reconstruct antibiotic resistance plasmids in <i>Escherichia coli</i>
97	<b>Chapter 4</b> Accurately reconstructing AMR plasmids of multiple bacterial species from short reads
143	<b>Chapter 5</b> Impact of selective digestive decontamination on the pangenome composition of ESBL- <i>E. coli</i>
175	<b>Chapter 6</b> Summary and General Discussion
187	<b>Appendices</b> English Summary Nederlandse samenvatting Acknowledgements About the Author

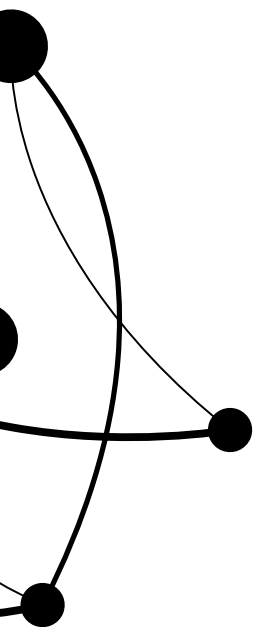




# 01

---

## General Introduction



### A brief overview of antimicrobial resistance

Antibiotics are regarded as one of the most important medical advances of the 20th century. Since the discovery of penicillin in 1928, antibiotics became the main tool to combat and prevent bacterial infections, saving the lives of many people and enabling significant progress in the field of medicine and surgery [1–3]. Furthermore, the use of antibiotics in multiple husbandry activities has contributed to the increase in production of high-quality food over the past 80 years [4]. Nevertheless, in 1940, soon after the discovery of penicillin, the first penicillin-resistant *Staphylococcus* strain was described [5]. Moreover, by the 1950s, penicillin resistance was already an important issue in clinical settings [6]. Since then, a myriad of new antimicrobials have been discovered, developed and deployed, and are of vital importance in the treatment of infections in ICU patients, among others. Unfortunately, bacteria eventually developed resistance mechanisms to counter every developed antimicrobial [2].

Resistance to antimicrobials can either be intrinsic or acquired. Intrinsic resistance occurs due to inherent properties of the bacterium [7], it does not depend on exposure to the antibiotic and it is generally encoded by chromosomally located features. For example, the presence of an outer membrane (OM), impermeable to many molecules, makes gram-negative bacteria (GNB) resistant to multiple antimicrobials [8]. In contrast, acquired resistance mechanisms may turn sensitive bacteria resistant to specific antimicrobials. There are several mechanisms by which bacteria may acquire resistance. The expression of efflux pumps can reduce the antimicrobial concentration to sub-inhibitory levels inside the cells. These pumps can either be non-specific, catalyzing the efflux of multiple drugs [9], or antibiotic-specific, such as tetracycline efflux pumps [10]. Additionally, changes in membrane permeability can reduce antimicrobial intake. Mutations in porins reduce the absorption rate of carbapenems in *Acinetobacter baumannii* and *Enterobacter cloacae* [11,12]. Antimicrobial resistance (AMR) can also arise due to the enzymatic degradation or modification of the antimicrobial. Resistance to multiple  $\beta$ -lactams is associated with the expression of bacterial enzymes, termed  $\beta$ -lactamases, that can hydrolyze this class of antibiotics rendering them ineffective [13]. Finally, bacteria can protect or modify the molecular target of the antibiotic. Vancomycin resistance in *Enterococcus* species arises due to alteration of the molecular target of vancomycin, i.e. the replacement of the d-Ala-d-Ala terminus of the peptidoglycan cell wall precursor lipid II, by d-Ala-d-Lac (VanA/VanB type of vancomycin resistance), thereby reducing the affinity between vancomycin and its target [14].

Although we may be tempted to think that antibiotic resistance is a novel phenomenon, it is in fact quite the opposite. Genes encoding resistance to  $\beta$ -lactam, tetracycline and glycopeptide antibiotics were found in ancient DNA from 30,000-year-old permafrost [15]. Moreover, bacterial strains detected in a region of the Lechuguilla Cave, New Mexico, that had been isolated for over 4 million years, were found to be resistant to 14 different commercially available antibiotics [16,17]. These findings conclusively show that antibiotic resistance is a natural phenomenon that predates the selective pressure of modern clinical antibiotic use.

### The rise of resistance

Today, AMR is considered one of the main threats to public health [18], with the number of infections caused by resistant bacteria consistently increasing each year [19]. In 2019, a total of 1.27 million deaths were attributed to bacterial AMR globally, with six pathogens being responsible for the majority: *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*,

*Acinetobacter baumannii* and *Pseudomonas aeruginosa*. Moreover, the emergence and global spread of multidrug resistant (MDR) bacteria, both in clinical and community settings, leaves physicians with few or no therapeutic options to treat infections caused by MDR bacteria. On top of this, only a limited number of new antimicrobials have been approved for human use in recent years [20].

The emergence and spread of AMR are complex phenomena, but are mainly driven by two factors. First, the use, overuse and misuse of antimicrobials in both humans and animals, which exerts an evolutionary pressure on bacteria and confers a selective survival advantage to those that have acquired resistance via mutations or more complex genomic rearrangements [21]. When bacteria replicate, these resistance traits are inherited (vertically) to the descendants. Within the human domain, the prevalence of antimicrobial resistant pathogens is highest in intensive care units (ICUs), where the most severely ill patients are admitted. They are often exposed to multiple antimicrobial drugs over extended periods of time. Second, bacteria can acquire exogenous DNA, including AMR genes, by horizontal gene transfer (HGT) [22]. The horizontal transfer of genes facilitates the spread of AMR determinants across bacterial populations in different niches.

The best described mechanisms of HGT in bacteria include transformation, transduction and conjugation. During transformation, bacteria take up free DNA from the environment, and incorporate it into their chromosome or reassemble it as part of the self-replicating episome [23]. Bacteria are capable of mediating the acquisition of large pieces of DNA (7 - 50 kb) through this mechanism [24]. Importantly, transformation has been shown to result in the transfer of clinically relevant AMR in a variety of human pathogens [25–28]. Transduction is mediated by phages and although different mechanisms for transduction exist, they all involve the mispackaging of bacterial DNA into the phage capsid. The phage then infects another host and transfers the genetic material to a different bacterial cell, where it will be integrated by homologous recombination [29,30]. Bacteriophages isolated from methicillin-resistant *Staphylococcus aureus* were found to transduce AMR genes to sensitive strains in the laboratory [31]. In gram-negative bacteria, transduction has been observed to transfer extended spectrum beta-lactamases (ESBL) genes from *Pseudomonas* hospital isolates to other *Pseudomonas* strains in-vitro [32] or a carbapenemase encoding gene was also transduced between *Acinetobacter* strains [33]. Conjugation requires cell-to-cell contact and the formation of a pore through which DNA can pass. The basic mechanism of conjugation is conserved among both gram-negative and gram-positive bacteria [34,35], and it was shown to occur in bacterial communities from diverse environments, such as soil, plant surfaces, sewage and even within the microbiome of animals [36–40]. Conjugative transfer systems are associated with plasmids and with integrative and conjugative elements (ICEs), also known as conjugative transposons. Plasmids are self-replicating extrachromosomal DNA elements that frequently carry AMR genes. From the three HGT mechanisms, conjugation is thought to be the most important quantitatively [41].

### **ESBL-producing *Escherichia coli***

*Escherichia coli* is a versatile micro-organism that can thrive in different ecological niches. It is a gram-negative facultative anaerobe bacterium that commonly resides as a commensal bacterium in the gut of humans and other warm-blooded animals [42,43]. However, several members of this species can also cause severe infections, either intra- or extra-intestinally [44]. The ‘success’ of *E. coli* as a pathogen is probably driven by its high degree of genomic plasticity. Gene loss events, DNA rearrangements, point mutations and the acquisition of genomic features via HGT have led to the

emergence of *E. coli* isolates that carry a wide repertoire of virulence factors, as well as multiple AMR genes [45–48]. Moreover, a recent assessment of the global burden of AMR estimated that resistant *E. coli* infections accounted for more than 250,000 deaths in 2019 [19].

In recent years, infections caused by extended spectrum beta-lactamase (ESBL)-producing *E. coli* have rapidly increased and have become an important public health concern. ESBLs are enzymes that can hydrolyze broad-spectrum  $\beta$ -lactams, including third generation cephalosporins, such as cefotaxime, ceftriaxone, and ceftazidime [49,50]. Different types of ESBL enzymes exist, including *bla*<sub>TEM</sub>, *bla*<sub>SHV</sub> and *bla*<sub>CTX-M</sub>. During the 1990s, *bla*<sub>TEM</sub> and *bla*<sub>SHV</sub> genes were dominant and mainly associated with specific clones of *K. pneumoniae*, generally isolated from clinical environments [51]. Nowadays, *E. coli* is the main ESBL-producing pathogen, being also the most prevalent resistant pathogen in Europe [48,52–54], and the second most prevalent worldwide [19], when considering both community and hospital-acquired infections. This increase in prevalence of ESBL-*E. coli* was triggered by the emergence and dissemination of the *bla*<sub>CTX-M</sub> genes, which is now the dominant ESBL gene family worldwide [51]. These genes are commonly associated with a wide variety of mobile genetic elements (MGE), like plasmids, ICEs and insertion sequences (IS), which facilitated their dissemination across the population structure of *E. coli* and partially explains the rapid increase in prevalence of ESBL-*E. coli* [51,55]. Additionally, the successful mutualistic association between IncF plasmids carrying the *bla*<sub>CTX-M-15</sub> or *bla*<sub>CTX-M-27</sub> variants with the globally disseminated clone ST131 further explains the success of this resistant pathogen [51,56,57]. *E. coli* ST131 appears to have an increased capability for gut colonisation and the existence of compensatory mutations in its chromosome seem to alleviate the fitness cost of carrying the aforementioned ESBL-plasmids [58,59].

It was observed that ESBL-*E. coli* frequently co-acquired resistance to fluoroquinolones and several other clinically important antimicrobials [60–62]. Therefore, carbapenems have become the first-choice antibiotic to treat invasive infections caused by ESBL-*E. coli*. However, the frequent use of carbapenems has led to an increased incidence of carbapenem resistant infections caused by Enterobacteriaceae, *Acinetobacter baumannii* and *Pseudomonas aeruginosa*, particularly in clinical environments [52,63].

### **ICU-acquired infections, colonisation and decontamination strategies in Dutch ICUs**

Healthcare-associated infections (HAIs) are those acquired by patients during their stay in a hospital or another healthcare setting. HAIs are associated with increased morbidity and mortality, and with an excess cost in patient care [64,65]. The European Center for Diseases prevention and Control estimated that 5.7% of patients admitted to a hospital in Europe in 2011 acquired a HAI [52]. Similar reports indicated that 4% of hospitalised patients in the U.S. acquired one of these infections in 2014 [66]. The most common pathogens found in healthcare associated infections are *E. coli*, *S. aureus*, *Enterococcus* spp., *Klebsiella* spp. and *P. aeruginosa* [52].

The prevalence of HAIs is the highest among patients admitted to ICUs, where approximately 1 in 5 patients (20%) will get a HAI during their stay [52]. Patients admitted to ICUs are generally at increased risk for acquiring infections, mainly due to severe disease and multiple comorbidities, interventions that reduce local defence mechanisms (e.g. mechanical ventilation, indwelling devices), as

well as, sometimes, the administration of immunosuppressive drugs, among other factors. These patients are susceptible to colonisation with potentially pathogenic microorganisms (PPMO), which is strongly associated with the development of infections in ICUs [67–70]. The gastrointestinal tract seems to be the primary reservoir for most bacterial pathogens associated with HAIs [71–75]. Once a patient has been colonised, cross-transmission of these pathogens across different patients is facilitated by their frequent contact with healthcare workers, abiotic surfaces or medical devices [76–83].

In the ICU, the prevalence of AMR is higher than in non-ICU wards. In the Netherlands, between 2015 and 2019, there was a slow increase in the prevalence of resistance to third-generation cephalosporins amongst *E. coli* diagnostic isolates in non-ICU wards from 5.5% to 6.5%, whereas in ICU wards, the prevalence increased from 8% to 10% [84].

Patients admitted to a Dutch ICU, if expected to undergo mechanical ventilation for at least 48 hours, receive selective digestive decontamination (SDD) as a prophylactic treatment to prevent colonisation with PPMOs. SDD consists of a mix of topical antibiotics (Tobramycin, Colistin and Amphotericin B) that aims to reduce the load of aerobic GNB, *P. aeruginosa*, *S. aureus* and yeast, but without compromising the anaerobic flora. SDD is administered as an oropharyngeal paste and as a solution through the nasogastric tube. Additionally, a 4-day course of an intravenous cephalosporin (cefotaxime or ceftriaxone) is also administered, to treat any incubating infection at the time of ICU admission [85]. A variation of SDD, named Selective Oropharyngeal Decontamination (SOD), consisting only of the oropharyngeal paste, is administered in some ICUs as an alternative to SDD [86]. In the Netherlands, where the prevalence of antibiotic resistance is low [87], SDD was associated with improved patient outcome in comparison to standard care, with reduced mortality, shorter lengths of ICU stay and a lower incidence of ICU-acquired bacteremia [88–93].

Multiple studies suggest that the use of SDD does not lead to an increase in the prevalence of colonisation or infection with antimicrobial resistant pathogens, however, those studies were based on phenotypic resistance data [94–96]. Nonetheless, studies that relied on metagenomics approaches reported important changes in the microbiome of patients treated with SDD, including an increase in the abundance of aminoglycoside-resistance genes [97,98]. It is important to note that those genes could not be linked to PPMOs and the studies were performed in small numbers of patients without a proper comparison group. The effect of SDD on the genome of gram-negative pathogens has not been studied before.

### **Plasmids, key players in the spread of AMR**

Plasmids are autonomously replicating DNA molecules that can coexist with the bacterial chromosome [99]. These genomic elements are ubiquitous in bacteria, and a single bacterial cell can harbour zero, one or multiple plasmids. Plasmids frequently mediate the transfer of beneficial accessory genes, such as AMR genes, within and between species of bacteria. These beneficial traits can confer resistance to antibiotics, expand the metabolic capabilities of bacteria and/or contribute to their adaptation to different environments [100–102]. Plasmids can be broadly categorised by their mobilization capabilities into conjugative, mobilizable and non-mobilizable [103]. Alternatively, they can also be categorized into incompatibility groups (Inc). Two plasmids belong to the same Inc type when they cannot stably coexist in the same bacterial cell.

Conjugative plasmids encode all necessary components to catalyze their own transmission to a different host via conjugation, and these plasmids have been associated with the spread of AMR since the late 1950s [104]. The genomic architecture and functionality of conjugative plasmids makes them ideal platforms for the dissemination of beneficial traits across bacterial populations. These are generally composed of a relatively stable plasmid backbone, which includes the genes encoding for conjugative capabilities but also for the systems that ensure the stability and vertical inheritance of the plasmid, such as replicative proteins, postsegregational killing systems and CRISPR-Cas arrays [105,106]. Aside from the backbone, conjugative plasmids usually contain a high density of repeated sequences, such as those associated with transposable elements, which creates hotspots for recombination. This proclivity towards recombination allows plasmids to easily capture blocks of genetic elements from distinct sources, but also to frequently rearrange or delete these blocks [107,108]. Consequently, plasmids often exhibit an extraordinary variety of accessory genes, arranged in mosaic-like structures [109–111]. This genetic plasticity also allows plasmids to accumulate multiple AMR genes in the same backbone, sometimes creating resistance islands, as observed in plasmids carrying IS26 in different gram-negative bacteria [112–114].

Studies describing plasmids as drivers behind the spread of AMR in different environments are accumulating. In the clinical environment, where the antibiotic pressure is high, AMR plasmids are increasingly being recognized as important contributors to the occurrence of prolonged single- and multi-species outbreaks [115–118]. Additionally, the dissemination of AMR plasmids, often carrying important resistance determinants as *bla*<sub>OXA-48</sub>, *bla*<sub>KPC-2</sub> and *bla*<sub>CTX-M-15</sub>, has been detected inside the gut of hospitalised patients [40,119–121]. Plasmids bearing multiple resistance genes with increased conjugative capabilities were also found in inhalable particulate matter of hospitals [122], which has serious implications for the potential sources of transmission of AMR in the clinics. In animal husbandry environments, where selective pressure is high, plasmids carrying resistance even to last-resource antibiotics were found are widespread in animals, soil and surrounding water sources [37,123–127]. More recently, the isolation of bacteria carrying plasmid-encoded carbapenemases or colistin resistance (*mcr*) genes in wild animals highlighted the high degree of AMR dissemination driven by plasmids [128,129].

### **Genomic surveillance of plasmids: Challenges and new methods**

Given that plasmids play an important role in the dissemination of AMR, it is becoming increasingly clear that if we want to better understand the mechanisms that drive this dissemination, we need tools that allow us to identify and classify plasmids in a fast, precise and high-throughput manner. The development of next-generation sequencing (NGS) platforms has allowed the characterization of bacterial genomes on a massive scale. Every platform works by initially fragmenting the genome and sequencing each fragment separately, producing reads. The subsequent *de novo* assembly of these reads leads to larger fragments of contiguous DNA sequence called contigs. NGS platforms can be classified according to the lengths of DNA reads they produce. Long reads, obtained via PacBio and Oxford Nanopore platforms, have an average length of around 20.000 bp [130,131], and after their *de novo* assembly it is possible to obtain complete genomes [132,133], meaning that each replicon in the bacteria (chromosome and plasmids) is represented by a single contig. The main disadvantages of long-read platforms are their higher costs and error rates. Short reads, obtained via Illumina sequencing platforms, can be up to 300 bp in length. Illumina sequencing platforms produce highly accurate reads and large amounts of samples can be processed simultaneously, making these platforms

cost-effective. Nevertheless, due to the frequent occurrence of repeated elements, the *de novo* assembly of short reads produces hundreds of contigs of unclear origin (plasmid or chromosome) mingled together in a draft genome. Consequently, determining the exact sequence of plasmids using short reads alone is challenging.

Although long reads allow obtaining complete genomes, Illumina short reads remain the most widely adopted sequencing technology in microbial genomics. As of July 2022, the sequence read archive (SRA) contained more than 1.8 million DNA sequences corresponding to bacterial genomes, and 98% of these were obtained using short reads (See Chapter 4 - Figure 1). Moreover, new short-read technologies are being brought to the market [134]. Consequently, there is interest in the development of tools that allow plasmid reconstruction from short-read data.

Multiple bioinformatic tools are currently available to predict bacterial plasmids from short-reads. They can be broadly categorised into two main classes:

- Binary classification tools use assembled contigs as input and classify them as plasmid- or chromosome-derived, therefore predicting the complete plasmid content of a bacterial strain, commonly known as the ‘plasmidome’, but without defining individual plasmids. These tools use a wide variety of computational approaches to classify contigs, including searches against databases of complete genomes [135,136], annotation of proteins [135,137] and machine learning or neural network classification algorithms [138].
- Plasmid reconstruction tools aim to predict individual plasmid sequences. To this end, tools can either use databases [139], information contained in the assembly graph in combination with coverage information [140], or both [141]. The output of these tools ideally permits studying the epidemiology of specific plasmids of interest [142].

A complete review of the different plasmid prediction tools can be found in chapter 2 of this thesis.

Although multiple tools exist to predict plasmids using short reads, a thorough and independent evaluation of tools performance across different species is missing. Consequently, choosing the best tool to perform prediction of plasmids (or plasmidome) is challenging.

### **Aim and outline of this thesis**

The aim of this thesis was to explore the differences in population structure, plasmidome and resistance compositions of a set of ESBL-*E. coli* isolates obtained from ICU-admitted patients that either received or did not receive SDD as a prophylactic treatment to prevent colonization with PPMOs, on top of standard care precautions. Given the difficulties for exploring the plasmidome using whole-genome sequencing (WGS) data, we first evaluated existing tools to predict plasmid sequences and subsequently developed new approaches to reconstruct plasmids using Illumina short reads.

In **chapter 2** we performed a comprehensive review of several tools designed to predict plasmids from short-read sequencing data. We also performed an exhaustive benchmark of six tools for the reconstruction of individual plasmids of 240 *E. coli* genomes. This study suggested that MOB-suite [139] and plasmidSPAdes [140] were the best performing tools. Nonetheless, we also discovered that all tools had major difficulties when reconstructing plasmids that contain AMR genes.

Consequently, in **chapter 3**, we developed a two-step method to improve the reconstruction of AMR plasmids of *E. coli* using short reads. In the first step, nodes in the assembly graph are classified as plasmid- or chromosome-derived by using plasmidEC, an ensemble classifier that we developed by combining three existing binary classification tools [135,137,143]. In the second step, we used gplas [144] to bin plasmid nodes into individual plasmid predictions based on similarities in sequence coverage and assembly graph connectivity. Gplas was also modified to better reconstruct plasmids that present large sequencing coverage variations. This method proved very successful for reconstructing AMR plasmids of *E. coli*, considerably outperforming MOB-suite in all evaluated metrics.

Therefore, in **chapter 4** we expanded our method to reconstruct plasmids of multiple species. To this end, we developed four species-specific models (for *E. faecium*, *K. pneumoniae*, *S. enterica*, *S. aureus*) and one species-independent plasmidEC model. By combining these models with gplas, our approach can be used to reconstruct plasmids of any species. To evaluate our tool, we reconstructed plasmids of more than 70 different species, and compared the performance against MOB-suite and plasmidSPAdes. We found that gplas performed consistently well when reconstructing large ARG-plasmids in multiple species, in contrast to a varying performance of the other tools.

In **chapter 5** we compared the pangenome, plasmidome and resistome composition of a set of ESBL-*E. coli* isolates from ICU patients that did or did not receive SDD. The data were derived from patients included in the R-GNOSIS ICU study [94], and encompassed isolates from five different ICUs located in Spain, Belgium and the UK. In this study we found that SDD had a limited impact on the population structure and pangenome composition of ESBL-*E. coli*. Nonetheless, isolates obtained from patients that received standard care had a higher amount of aminoglycoside resistance genes, while SDD isolates were more frequently found to possess a transposon carrying a tobramycin resistance gene. This transposon contained a total of 3 ARG genes surrounded by IS26 elements, and frequently co-occurred with *bla*<sub>CTX-M-15</sub> in multiple clones and distinct plasmid backbones.

Finally, in **chapter 6**, I discuss the new methods developed in this thesis, their main limitations and their potential applications. Moreover, I describe and reflect on the main findings on the population genomics of *E. coli* recovered from ICU patients in the R-GNOSIS ICU study, and formulate recommendations for future research.



## References

1. Lobanovska M, Pilla G. Focus: Drug Development: Penicillin's Discovery and Antibiotic Resistance: Lessons for the Future? *Yale J Biol Med.* 2017;90: 135.
2. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. *P T.* 2015;40. Available: <https://pubmed.ncbi.nlm.nih.gov/25859123/>
3. Adedeji WA. The treasure called antibiotics. *Annals of Ibadan postgraduate medicine.* 2016;14. Available: <https://pubmed.ncbi.nlm.nih.gov/28337088/>
4. Kirchhelle C. *Pharming animals: a global history of antibiotics in food production (1935–2017).* Palgrave Communications. 2018. doi:10.1057/s41599-018-0152-2
5. Spink WW, Ferris V. Penicillin-resistant *Staphylococci*: mechanisms involved in the development of resistance. *J Clin Invest.* 1947;26: 379–393.
6. Spellberg B, Gilbert DN. The Future of Antibiotics and Resistance: A Tribute to a Career of Leadership by John Bartlett. *Clinical Infectious Diseases.* 2014. pp. S71–S75. doi:10.1093/cid/ciu392
7. Cox G, Wright GD. Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *Int J Med Microbiol.* 2013;303: 287–292.
8. Nikaido H. Prevention of drug access to bacterial targets: permeability barriers and active efflux. *Science.* 1994;264: 382–388.
9. Vargiu AV, Pos KM, Poole K, Nikaido H. Bad Bugs in the XXIst Century: Resistance Mediated by Multi-Drug Efflux Pumps in Gram-Negative Bacteria. *Frontiers Media SA;* 2016.
10. Roberts MC. Tetracycline resistance determinants: mechanisms of action, regulation of expression, genetic mobility, and distribution. *FEMS Microbiol Rev.* 1996;19: 1–24.
11. Uppalapati SR, Sett A, Pathania R. The Outer Membrane Proteins OmpA, CarO, and OprD of *Acinetobacter baumannii* Confer a Two-Pronged Defense in Facilitating Its Success as a Potent Human Pathogen. *Front Microbiol.* 2020;11. doi:10.3389/fmicb.2020.589234
12. Babouee FB, Ellington MJ, Hopkins KL, Turton JF, Doumith M, Loy R, et al. Association of Novel Nonsynonymous Single Nucleotide Polymorphisms in ampD with Cephalosporin Resistance and Phylogenetic Variations in ampC, ampR, ompF, and ompC in *Enterobacter cloacae* Isolates That Are Highly Resistant to Carbapenems. *Antimicrob Agents Chemother.* 2016;60. doi:10.1128/AAC.02835-15
13. Bush K. Bench-to-bedside review: The role of beta-lactamases in antibiotic-resistant Gram-negative infections. *Crit Care.* 2010;14: 224.
14. Stogios PJ, Savchenko A. Molecular mechanisms of vancomycin resistance. *Protein Sci.* 2020;29: 654–669.
15. D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, et al. Antibiotic resistance is ancient. *Nature.* 2011;477. doi:10.1038/nature10388
16. Bhullar K, Waglechner N, Pawlowski A, Koteva K, Banks ED, Johnston, et al. Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS One.* 2012;7. doi:10.1371/journal.pone.0034953
17. Pawlowski AC, Wang W, Koteva K, Barton HA, McArthur AG, Wright GD. A diverse intrinsic antibiotic resistome from a cave bacterium. *Nat Commun.* 2016;7. doi:10.1038/ncomms13803
18. Global action plan on antimicrobial resistance. World Health Organization; 1 Jan 2016 [cited 8 Mar 2023]. Available: <https://www.who.int/publications/i/item/9789241509763>
19. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet.* 2022;399: 629–655.
20. Terreni M, Taccani M, Pregnolato M. New Antibiotics for Multidrug-Resistant Bacterial Strains: Latest Research Developments and Future Perspectives. *Molecules.* 2021;26: 2671.

21. Oz T, Guvenek A, Yildiz S, Karaboga E, Tamer YT, Mumcuyan N, et al. Strength of Selection Pressure Is an Important Parameter Contributing to the Complexity of Antibiotic Resistance Evolution. *Molecular Biology and Evolution*. 2014. pp. 2387–2401. doi:10.1093/molbev/msu191
22. Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. *Virulence Mechanisms of Bacterial Pathogens*. 2016. pp. 481–511. doi:10.1128/9781555819286.ch17
23. Natural competence for transformation. *Curr Biol*. 2016;26: R1126–R1130.
24. In and out contribution of natural transformation to the shuffling of large genomic regions. *Curr Opin Microbiol*. 2017;38: 22–29.
25. Lu J, Wang Y, Zhang S, Bond P, Yuan Z, Guo J. Triclosan at environmental concentrations can enhance the spread of extracellular antibiotic resistance genes through transformation. *Sci Total Environ*. 2020;713: 136621.
26. Traglia GM, Place K, Dotto C, Fernandez JS, Montaña S, Bahiense CDS, et al. Interspecies DNA acquisition by a naturally competent *Acinetobacter baumannii* strain. *Int J Antimicrob Agents*. 2019;53: 483–490.
27. Alexander HE, Hahn E, Leidy G. On the specificity of the desoxyribonucleic acid which induces streptomycin resistance in *Hemophilus*. *J Exp Med*. 1956;104: 305–320.
28. Janoir C, Podglajen I, Kitzis M, Poyart C, Gutmann L. In Vitro Exchange of Fluoroquinolone Resistance Determinants between *Streptococcus pneumoniae* and Viridans Streptococci and Genomic Organization of the *parE-parC* region in *S. mitis*. *The Journal of Infectious Diseases*. 1999. pp. 555–558. doi:10.1086/314888
29. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol*. 2015;13: 641–650.
30. Humphrey S, Fillol-Salom A, Quiles-Puchalt N, Ibarra-Chávez R, Haag AF, Chen J, et al. Bacterial chromosomal mobility via lateral transduction exceeds that of classical mobile genetic elements. *Nat Commun*. 2021;12: 6509.
31. Stanczak-Mrozek KI, Manne A, Knight GM, Gould K, Witney AA, Lindsay JA. Within-host diversity of MRSA antimicrobial resistances. *J Antimicrob Chemother*. 2015;70: 2191–2198.
32. Blahová J, Králiková K, Krcmery V, Jezek P. Low-Frequency transduction of imipenem resistance and high-frequency transduction of ceftazidime and aztreonam resistance by the bacteriophage AP-151 isolated from a *Pseudomonas aeruginosa* strain. *J Chemother*. 2000;12: 482–486.
33. Krahn T, Wibberg D, Maus I, Winkler A, Bontron S, Sczyrba A, et al. Intraspecies Transfer of the Chromosomal *Acinetobacter baumannii* blaNDM-1 Carbapenemase Gene. *Antimicrobial Agents and Chemotherapy*. 2016. pp. 3032–3040. doi:10.1128/aac.00124-16
34. Grohmann E, Muth G, Espinosa M. Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev*. 2003;67: 277–301, table of contents.
35. de la Cruz F, Frost LS, Meyer RJ, Zechner EL. Conjugative DNA metabolism in Gram-negative bacteria. *FEMS Microbiol Rev*. 2010;34: 18–40.
36. Davison J. Genetic exchange between bacteria in the environment. *Plasmid*. 1999;42: 73–91.
37. Meng M, Li Y, Yao H. Plasmid-Mediated Transfer of Antibiotic Resistance Genes in Soil. *Antibiotics*. 2022;11. doi:10.3390/antibiotics11040525
38. Li C, Chen J, Li SC. Understanding Horizontal Gene Transfer network in human gut microbiota. *Gut Pathog*. 2020;12: 1–20.
39. Lerner A, Matthias T, Aminov R. Potential Effects of Horizontal Gene Exchange in the Human Gut. *Front Immunol*. 2017;8: 1630.
40. León-Sampedro R, DelaFuente J, Díaz-Agero C, Crellen T, Musicha P, Rodríguez-Beltrán J, et al. Pervasive transmission of a carbapenem resistance plasmid in the gut microbiota of hospitalized patients. *Nature microbiology*. 2021;6. doi:10.1038/s41564-021-00879-y
41. Halary S, Leigh JW, Cheaib B, Lopez P, Baptiste E. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences*. 2010. pp. 127–132. doi:10.1073/pnas.0908978107

42. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486: 207–214.
43. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. *Nat Rev Microbiol*. 2004;2. doi:10.1038/nrmicro818
44. Biran D, Ron EZ. Extraintestinal Pathogenic *Escherichia coli*. *Current Topics in Microbiology and Immunology*. 2018. pp. 149–161. doi:10.1007/82\_2018\_108
45. Leimbach A, Hacker J, Dobrindt U. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Curr Top Microbiol Immunol*. 2013;358: 3–32.
46. Desvaux M, Dalmasso G, Beyrouthy R, Barnich N, Delmas J, Bonnet R. Pathogenicity Factors of Genomic Islands in Intestinal and Extraintestinal. *Front Microbiol*. 2020;11: 2065.
47. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nature Reviews Microbiology*. 2021. pp. 37–54. doi:10.1038/s41579-020-0416-x
48. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis*. 2019;19: 56–66.
49. Jacoby GA, Carreras I. Activities of beta-lactam antibiotics against *Escherichia coli* strains producing extended-spectrum beta-lactamases. *Antimicrob Agents Chemother*. 1990;34: 858–862.
50. Farber B, Moellering RC Jr. The third generation cephalosporins. *Bull N Y Acad Med*. 1982;58: 696–710.
51. Castanheira M, Simmer PJ, Bradford PA. Extended-spectrum  $\beta$ -lactamases: an update on their characteristics, epidemiology and detection. *JAC Antimicrob Resist*. 2021;3: dlab092.
52. European Centre for Disease Prevention and Control. Point prevalence survey of healthcare associated infections and antimicrobial use in European acute care hospitals. Stockholm: ECDC; 2013.
53. European Centre for Disease Prevention and Control. Healthcare-associated infections in intensive care units - Annual Epidemiological Report for 2017. ECDC; 2019.
54. European Centre for Disease Prevention and Control. Assessing the health burden of infections with antibiotic-resistant bacteria in the EU/EEA, 2016-2020.
55. Kawamura K, Nagano N, Suzuki M, Wachino J-I, Kimura K, Arakawa Y. ESBL-producing *Escherichia coli* and Its Rapid Rise among Healthy People. *Food Safety*. 2017;5: 122.
56. Nicolas-Chanoine MH, Bertrand X, Madec JY. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev*. 2014;27. doi:10.1128/CMR.00125-13
57. Pitout JD, DeVinney R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Res*. 2017;6. doi:10.12688/f1000research.10609.1
58. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, et al. Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection. *mBio*. 2019. doi:10.1128/mbio.00644-19
59. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet*. 2016;12: e1006280.
60. Cerquetti M, Giufrè M, García-Fernández A, Accogli M, Fortini D, Luzzi I, et al. Ciprofloxacin-resistant, CTX-M-15-producing *Escherichia coli* ST131 clone in extraintestinal infections in Italy. *Clin Microbiol Infect*. 2010;16: 1555–1558.
61. Blanco J, Mora A, Mamani R, López C, Blanco M, Dahbi G, et al. National survey of *Escherichia coli* causing extraintestinal infections reveals the spread of drug-resistant clonal groups O25b:H4-B2-ST131, O15:H1-D-ST393 and CGA-D-ST69 with high virulence gene content in Spain. *J Antimicrob Chemother*. 2011;66: 2011–2021.

62. Kim S-Y, Park Y-J, Johnson JR, Yu JK, Kim Y-K, Kim YS. Prevalence and characteristics of *Escherichia coli* sequence type 131 and its H30 and H30Rx subclones: a multicenter study from Korea. *Diagn Microbiol Infect Dis.* 2016;84: 97–101.
63. European Centre for Disease Prevention and Control. Healthcare-associated infections acquired in intensive care units. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019.
64. Allegranzi B. Report on the Burden of Endemic Health Care-Associated Infection Worldwide. Sudan R, editor. World Health Organization;
65. Umscheid CA, Mitchell, Doshi JA, Agarwal R, Williams K, Brennan PJ. Estimating the proportion of healthcare-associated infections that are reasonably preventable and the related mortality and costs. *Infect Control Hosp Epidemiol.* 2011;32. doi:10.1086/657912
66. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, et al. Multistate Point-Prevalence Survey of Health Care–Associated Infections. *N Engl J Med.* 2014;370: 1198.
67. Healthcare-associated infections in adult intensive care unit patients: Changes in epidemiology, diagnosis, prevention and contributions of new technologies. *Intensive Crit Care Nurs.* 2022;70: 103227.
68. Ziakas PD, Thapa R, Rice LB, Mylonakis E. Trends and significance of VRE colonization in the ICU: a meta-analysis of published studies. *PLoS One.* 2013;8: e75658.
69. Garrouste-Orgeas M, Timsit J-F, Kallel H, Ben Ali A, Dumay MF, Paoli B, et al. Colonization With Methicillin-Resistant *Staphylococcus aureus* in ICU Patients Morbidity, Mortality, and Glycopeptide Use. *Infection Control & Hospital Epidemiology.* 2001. pp. 687–692. doi:10.1086/501846
70. Frencken JF, Wittekamp BHJ, Plantinga NL, Spitoni C, van de Groep K, Cremer OL, et al. Associations Between Enteral Colonization With Gram-Negative Bacteria and Intensive Care Unit–Acquired Infections and Colonization of the Respiratory Tract. *Clinical Infectious Diseases.* 2018. pp. 497–503. doi:10.1093/cid/cix824
71. Sommerstein R, Merz TM, Berger S, Kraemer JG, Marschall J, Hilty M. Patterns in the longitudinal oropharyngeal microbiome evolution related to ventilator-associated pneumonia. *Antimicrobial Resistance & Infection Control.* 2019. doi:10.1186/s13756-019-0530-6
72. Vaishnavi C. Translocation of gut flora and its role in sepsis. *Indian Journal of Medical Microbiology.* 2013. pp. 334–342. doi:10.4103/0255-0857.118870
73. Berg RD. Bacterial translocation from the gastrointestinal tract. *Adv Exp Med Biol.* 1999;473. doi:10.1007/978-1-4615-4143-1\_2
74. DeFilipp Z, Bloom PP, Torres Soto M, Mansour MK, Sater MRA, Huntley MH, et al. Drug-Resistant Bacteremia Transmitted by Fecal Microbiota Transplant. *N Engl J Med.* 2019;381: 2043–2050.
75. Karanika S, Karantanos T, Arvanitis M, Grigoras C, Mylonakis E. Fecal Colonization With Extended-spectrum Beta-lactamase–Producing Enterobacteriaceae and Risk Factors Among Healthy Individuals: A Systematic Review and Metaanalysis. *Clin Infect Dis.* 2016;63: 310–318.
76. Sommerstein R, Merz TM, Berger S, Kraemer JG, Marschall J, Hilty M. Patterns in the longitudinal oropharyngeal microbiome evolution related to ventilator-associated pneumonia. *Antimicrobial Resistance & Infection Control.* 2019. doi:10.1186/s13756-019-0530-6
77. Vaishnavi C. Translocation of gut flora and its role in sepsis. *Indian Journal of Medical Microbiology.* 2013. pp. 334–342. doi:10.4103/0255-0857.118870
78. Berg RD. Bacterial translocation from the gastrointestinal tract. *Adv Exp Med Biol.* 1999;473. doi:10.1007/978-1-4615-4143-1\_2
79. DeFilipp Z, Bloom PP, Torres Soto M, Mansour MK, Sater MRA, Huntley MH, et al. Drug-Resistant Bacteremia Transmitted by Fecal Microbiota Transplant. *N Engl J Med.* 2019;381: 2043–2050.
80. Karanika S, Karantanos T, Arvanitis M, Grigoras C, Mylonakis E. Fecal Colonization With Extended-spectrum Beta-lactamase–Producing Enterobacteriaceae and Risk Factors Among Healthy Individuals: A

Systematic Review and Metaanalysis. *Clin Infect Dis*. 2016;63: 310–318.

81. Hu Y, Zhang H, Wei L, Feng Y, Wen H, Li J, et al. Competitive Transmission of Carbapenem-Resistant *Klebsiella pneumoniae* in a Newly Opened Intensive Care Unit. *mSystems*. 2022;7. doi:10.1128/msystems.00799-22
82. Yan Z, Zhou Y, Du M, Bai Y, Liu B, Gong M, et al. Prospective investigation of carbapenem-resistant *Klebsiella pneumoniae* transmission among the staff, environment and patients in five major intensive care units, Beijing. *J Hosp Infect*. 2019;101. doi:10.1016/j.jhin.2018.11.019
83. Austin DJ, Bonten MJM, Weinstein RA, Slaughter S, Anderson RM. Vancomycin-resistant enterococci in intensive-care hospital settings: Transmission dynamics, persistence, and the impact of infection control programs. *Proceedings of the National Academy of Sciences*. 1999. pp. 6908–6913. doi:10.1073/pnas.96.12.6908
84. NethMap 2022. Consumption of antimicrobial agents and antimicrobial resistance among medically important bacteria in the Netherlands. [cited 20 Apr 2023]. Available: <https://swab.nl/en/nethmap-pvid369?>
85. Wittekamp BHJ, Oostdijk EAN, Cuthbertson BH, Brun-Buisson C, Bonten MJM. Selective decontamination of the digestive tract (SDD) in critically ill patients: a narrative review. *Intensive Care Med*. 2019;46: 343–349.
86. Bergmans DCJJ, Bergmans DCJ, Bonten MJM, Gaillard CA, Paling JC, van der GEEST S, et al. Prevention of Ventilator-associated Pneumonia by Oral Decontamination. *American Journal of Respiratory and Critical Care Medicine*. 2001. pp. 382–388. doi:10.1164/ajrccm.164.3.2005003
87. WHO Regional Office for Europe/European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2022 – 2020 data. Copenhagen: WHO Regional Office for Europe; 2022. 2022.
88. Jonge E de, de Jonge E, Schultz MJ, Spanjaard L, Bossuyt PMM, Vroom MB, et al. Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial. *The Lancet*. 2003. pp. 1011–1016. doi:10.1016/s0140-6736(03)14409-1
89. Jonge E de, de Jonge E. Effects of selective decontamination of digestive tract on mortality and antibiotic resistance in the intensive-care unit. *Current Opinion in Critical Care*. 2005. pp. 144–149. doi:10.1097/01.ccx.0000155352.01489.11
90. Oostdijk EAN, Kesecioglu J, Schultz MJ, Visser CE, de Jonge E, van Essen EHR, et al. Effects of decontamination of the oropharynx and intestinal tract on antibiotic resistance in ICUs: a randomized clinical trial. *JAMA*. 2014;312: 1429–1437.
91. van Hout D, Plantinga NL, Bruijning-Verhagen PC, Oostdijk EAN, de Smet AMGA, de Wit GA de, et al. Cost-effectiveness of selective digestive decontamination (SDD) versus selective oropharyngeal decontamination (SOD) in intensive care units with low levels of antimicrobial resistance: an individual patient data meta-analysis. *BMJ Open*. 2019;9: e028876.
92. de Smet AM, Kluytmans JA, Cooper BS, Mascini EM, Benus RF, van der Werf TS, et al. Decontamination of the digestive tract and oropharynx in ICU patients. *N Engl J Med*. 2009;360. doi:10.1056/NEJMoa0800394
93. Houben AJM, Oostdijk EAN, van der Voort PHJ, Monen JCM, Bonten MJM, van der Bij AK, et al. Selective decontamination of the oropharynx and the digestive tract, and antimicrobial resistance: a 4 year ecological study in 38 intensive care units in the Netherlands. *J Antimicrob Chemother*. 2014;69: 797–804.
94. Wittekamp BH, Plantinga NL, Cooper BS, Lopez-Contreras J, Coll P, Mancebo J, et al. Decontamination Strategies and Bloodstream Infections With Antibiotic-Resistant Microorganisms in Ventilated Patients: A Randomized Clinical Trial. *JAMA*. 2018;320: 2087–2098.
95. Daneman N, Sarwar S, Fowler RA, Cuthbertson BH, SuDDICU Canadian Study Group. Effect of selective decontamination on antimicrobial resistance in intensive care units: a systematic review and meta-analysis. *Lancet Infect Dis*. 2013;13: 328–341.

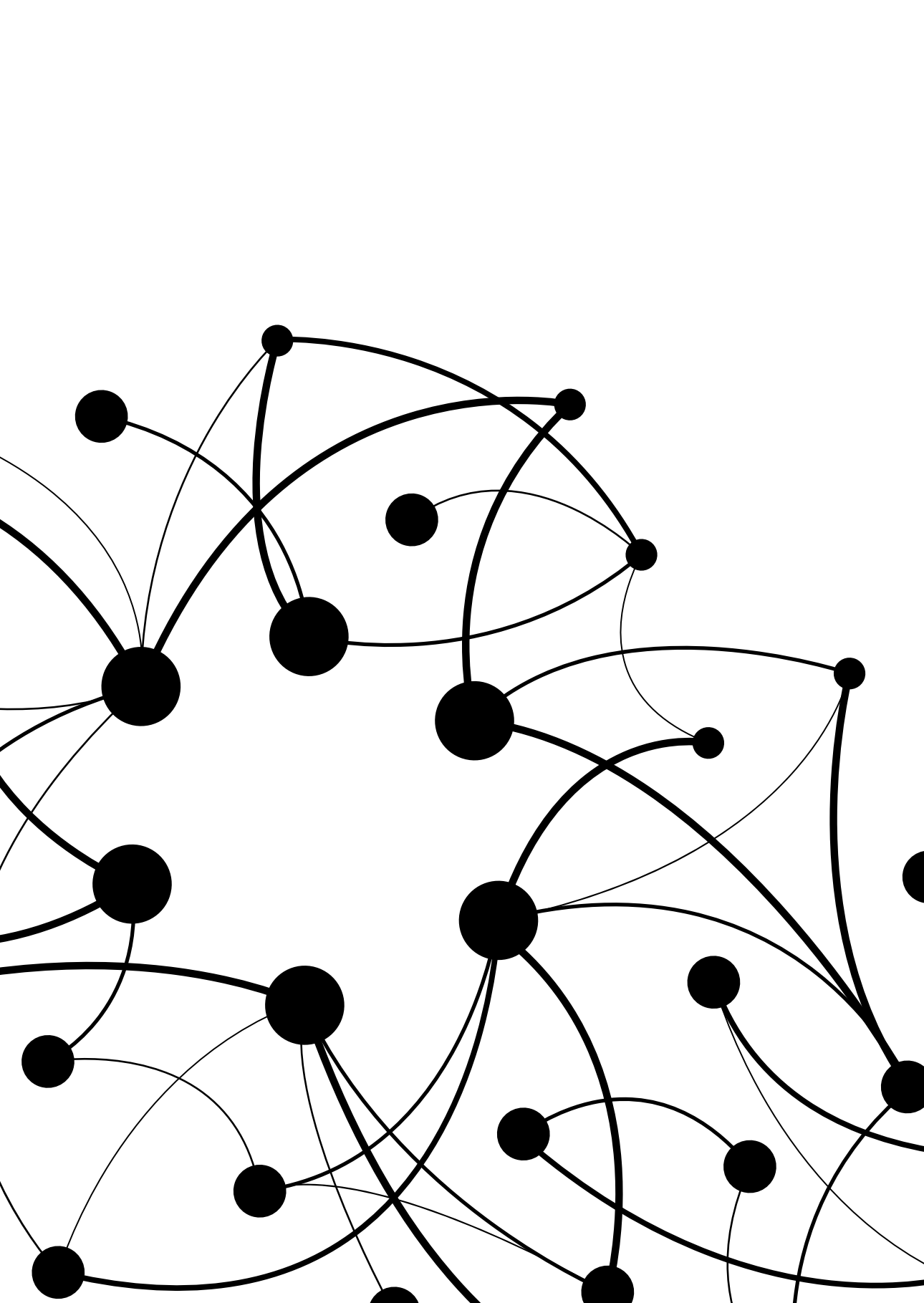
96. Bonten MJM, Oostdijk EAN, van der Bij AK. Selective decontamination of the oropharynx and the digestive tract, and antimicrobial resistance: a 4 year ecological study in 38 intensive care units in the Netherlands--authors' response. *The Journal of antimicrobial chemotherapy*. 2014. p. 861.
97. Buelow E, Bello González TDJ, Fuentes S, de Steenhuijsen Piters WAA, Lahti L, Bayjanov JR, et al. Comparative gut microbiota and resistome profiling of intensive care patients receiving selective digestive tract decontamination and healthy subjects. *Microbiome*. 2017;5: 88.
98. Buelow E, Gonzalez TB, Versluis D, Oostdijk EAN, Ogilvie LA, van Mourik MSM, et al. Effects of selective digestive decontamination (SDD) on the gut resistome. *J Antimicrob Chemother*. 2014;69: 2215–2223.
99. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat Rev Microbiol*. 2021;19: 347–359.
100. Carattoli A. Resistance Plasmid Families in *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy*. 2009. pp. 2227–2238. doi:10.1128/aac.01707-08
101. Palomino A, Gewurz D, DeVine L, Zajmi U, Morales J, Abu-Rumman F, et al. Metabolic genes on conjugative plasmids are highly prevalent in *Escherichia coli* and can protect against antibiotic treatment. *ISME J*. 2023;17: 151–162.
102. Pilla G, Tang CM. Going around in circles: virulence plasmids in enteric pathogens. *Nat Rev Microbiol*. 2018;16: 484–495.
103. Smillie C, Pilar Garcillán-Barcia M, Victoria Francia M, Rocha EPC, de la Cruz F. Mobility of Plasmids. *Microbiol Mol Biol Rev*. 2010;74: 434.
104. Watanabe T. Infective heredity of multiple drug resistance bacteria. *Bacteriological Reviews*. 1963. pp. 87–115. doi:10.1128/br.27.1.87-115.1963
105. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol*. 2020;18. doi:10.1038/s41579-019-0278-2
106. Fernández-López R, Garcillán-Barcia MP, Revilla C, Lázaro M, Vielva L, de la Cruz F. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol Rev*. 2006;30: 942–966.
107. Rodríguez-Beltrán J, Turret J, Tenaillon O, López E, Bourdelier E, Costas C, et al. High Recombinant Frequency in Extraintestinal Pathogenic *Escherichia coli* Strains. *Mol Biol Evol*. 2015;32: 1708–1716.
108. Niaudet B, Jannière L, Ehrlich SD. Recombination between repeated DNA sequences occurs more often in plasmids than in the chromosome of *Bacillus subtilis*. *Mol Gen Genet*. 1984;197: 46–54.
109. Toussaint A, Merlin C. Mobile elements as a combination of functional modules. *Plasmid*. 2002;47: 26–35.
110. Boyd EF, Fidelma Boyd E, Hill CW, Rich SM, Hartl DL. Mosaic Structure of Plasmids From Natural Populations of *Escherichia coli*. *Genetics*. 1996. pp. 1091–1100. doi:10.1093/genetics/143.3.1091
111. Pesesky MW, Tilley R, Beck DAC. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid*. 2019;102: 10–18.
112. Varani A, He S, Siguier P, Ross K, Chandler M. The IS6 family, a clinically important group of insertion sequences including IS26. *Mobile DNA*. 2021. doi:10.1186/s13100-021-00239-x
113. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev*. 2018;31. doi:10.1128/CMR.00088-17
114. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, 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. doi:10.1073/pnas.2008731118
115. Yamagishi T, Matsui M, Sekizuka T, Ito H, Fukusumi M, Uehira T, et al. A prolonged multispecies outbreak of IMP-6 carbapenemase-producing Enterobacterales due to horizontal transmission of the IncN plasmid. *Sci Rep*. 2020;10: 4139.

116. Mari-Almirall M, Ferrando N, Fernández MJ, Cosgaya C, Viñes J, Rubio E, et al. Clonal Spread and Intra- and Inter-Species Plasmid Dissemination Associated With *Klebsiella pneumoniae* Carbapenemase-Producing Enterobacterales During a Hospital Outbreak in Barcelona, Spain. *Front Microbiol.* 2021;12. doi:10.3389/fmicb.2021.781127
117. Hidalgo L, de Been M, Rogers MRC, Schürch AC, Scharringa J, van der Zee A, et al. Sequence-based epidemiology of an OXA-48 plasmid during a hospital outbreak. *Antimicrob Agents Chemother.* 2019;63. doi:10.1128/AAC.01204-19
118. Hawkey J, Wyres KL, Judd LM, Harshegyi T, Blakeway L, Wick RR, et al. ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Med.* 2022;14: 1–13.
119. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. 2020 [cited 10 Mar 2023]. doi:10.7554/eLife.53886
120. Schweizer C, Bischoff P, Bender J, Kola A, Gastmeier P, Hummel M, et al. Plasmid-Mediated Transmission of KPC-2 Carbapenemase in Critically Ill Patients. *Front Microbiol.* 2019;10: 276.
121. Ruan Z, Feng Y, Tang Y-W, Wang S, Uhlemann A-C. New Insights Into the Transmission Dynamics and Control of Antimicrobial Resistance to Last-resort Antibiotics. *Frontiers Media SA;* 2022.
122. Zhou Z-C, Shuai X-Y, Lin Z-J, Liu Y, Zhu L, Chen H. Prevalence of multi-resistant plasmids in hospital inhalable particulate matter (PM) and its impact on horizontal gene transfer. *Environ Pollut.* 2021;270: 116296.
123. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, et al. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother.* 2018;73: 1121–1137.
124. Lihan S, Lee SY, Toh SC, Leong SS. Plasmid-Mediated Antibiotic Resistant *Escherichia coli* in Sarawak Rivers and Aquaculture Farms, Northwest of Borneo. *Antibiotics.* 2021;10. doi:10.3390/antibiotics10070776
125. Checucci A, Trevisi P, Luise D, Modesto M, Blasioli S, Braschi I, et al. Exploring the Animal Waste Resistome: The Spread of Antimicrobial Resistance Genes Through the Use of Livestock Manure. *Front Microbiol.* 2020;11. doi:10.3389/fmicb.2020.01416
126. Nesporova K, Valcek A, Papagiannitsis C, Kutilova I, Jamborova I, Davidova-Gerzova L, et al. Multi-Drug Resistant Plasmids with ESBL/AmpC and mcr-5.1 in Paraguayan Poultry Farms: The Linkage of Antibiotic Resistance and Hatcheries. *Microorganisms.* 2021;9: 866.
127. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci Adv.* 2021;7. doi:10.1126/sciadv.abe3868
128. Wang J, Ma Z-B, Zeng Z-L, Yang X-W, Huang Y, Liu J-H. The role of wildlife (wild birds) in the global transmission of antimicrobial resistance genes. *Zool Res.* 2017;38: 55–80.
129. Dolejska M, Papagiannitsis CC. Plasmid-mediated resistance is going wild. *Plasmid.* 2018;99: 99–111.
130. Arakawa K. Nanopore Sequencing: Methods and Protocols. Springer Nature; 2023.
131. Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data.* 2020;7: 399.
132. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics.* 2017;3. doi:10.1099/mgen.0.000132
133. Arredondo-Alonso S, Pöntinen AK, Cléon F, Gladstone RA, Schürch AC, Johnsen PJ, et al. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience.* 2021;10. doi:10.1093/gigascience/giab079

- 134.** Eisenstein M. Innovative technologies crowd the short-read sequencing market. *Nature*. 2023;614: 798–800.
- 135.** Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000211
- 136.** Gomi R, Wyres KL, Holt KE. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb Genom*. 2021;7. doi:10.1099/mgen.0.000550
- 137.** Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom*. 2020;6. doi:10.1099/mgen.0.000398
- 138.** Pu L, Shamir R. 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics*. 2022;38: ii56–ii61.
- 139.** Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial genomics*. 2018;4. doi:10.1099/mgen.0.000206
- 140.** Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016;32: 3380–3387.
- 141.** Müller R, Chauve C. HyAsP, a greedy tool for plasmids identification. *Bioinformatics*. 2019;35: 4436–4439.
- 142.** Arredondo-Alonso S, Top J, Corander J, Willems RJL, Schürch AC. Mode and dynamics of vanA-type vancomycin resistance dissemination in Dutch hospitals. *Genome Med*. 2021;13: 9.
- 143.** 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. doi:10.1099/mgen.0.000683
- 144.** Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics*. 2020;36: 3874–3876.







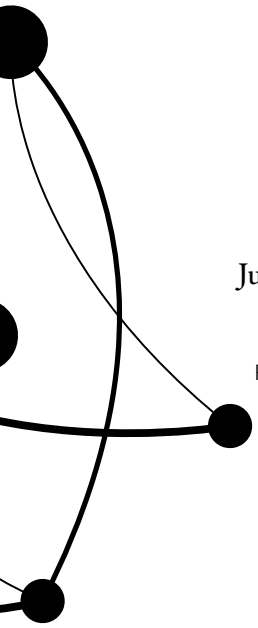
# 02

---

## Recovering *E. coli* plasmids in the absence of long-read sequencing data

Julian A. Paganini, Nienke L. Plantinga, Sergio Arredondo-Alonso,  
Rob J. L. Willems and Anita C. Schürch

Published in MDPI *microorganisms*. (2021). doi: 10.3390/microorganisms9081613



### Abstract

The incidence of infections caused by multidrug-resistant *E. coli* strains has risen in the past years. Antibiotic resistance in *E. coli* is often mediated by acquisition and maintenance of plasmids. The study of *E. coli* plasmid epidemiology and genomics often requires long-read sequencing information, but recently a number of tools that allow plasmid prediction from short-read data have been developed. Here, we reviewed 25 available plasmid prediction tools and categorized them into binary plasmid/chromosome classification tools and plasmid reconstruction tools. We benchmarked six tools (MOB-suite, plasmidSPAdes, gplas, FishingForPlasmids, HyAsP and SCAPP) that aim to reliably reconstruct distinct plasmids, with a special focus on plasmids carrying antibiotic resistance genes (ARGs) such as extended-spectrum beta-lactamase genes. We found that two thirds ( $n = 425$ , 66.3%) of all plasmids were correctly reconstructed by at least one of the six tools, with a range of 92 (14.58%) to 317 (50.23%) correctly predicted plasmids. However, the majority of plasmids that carried antibiotic resistance genes ( $n = 85$ , 57.8%) could not be completely recovered as distinct plasmids by any of the tools. MOB-suite was the only tool that was able to correctly reconstruct the majority of plasmids ( $n = 317$ , 50.23%), and performed best at reconstructing large plasmids ( $n = 166$ , 46.37%) and ARG-plasmids ( $n = 41$ , 27.9%), but predictions frequently contained chromosome contamination (40%). In contrast, plasmidSPAdes reconstructed the highest fraction of plasmids smaller than 18 kbp ( $n = 168$ , 61.54%). Large ARG-plasmids, however, were frequently merged with sequences derived from distinct replicons. Available bioinformatic tools can provide valuable insight into *E. coli* plasmids, but also have important limitations. This work will serve as a guideline for selecting the most appropriate plasmid reconstruction tool for studies focusing on *E. coli* plasmids in the absence of long-read sequencing data.

## Introduction

*Escherichia coli* is a versatile micro-organism able to survive and thrive in different ecological habitats. It is a Gram-negative facultative anaerobe that commonly resides in the human gut as a commensal bacteria [1]. However, several members of this species also harbor the potential to cause severe infections, both intestinally [2] and extra-intestinally [3], in the healthcare settings [4] as well as in the community [5]. The ‘success’ of *E. coli* as a pathogen can be mostly attributed to the wide repertoire of virulence factors that strains may carry [6] and the increasing fraction of infections caused by multidrug-resistant strains [7]. Many of the antibiotic resistance genes and virulence factors present in *E. coli* are commonly encoded on plasmids, mobile genetic elements (MGE) that can be horizontally disseminated [8,9,10]. Therefore, precise identification and characterization of *E. coli* plasmids are highly relevant from an epidemiological and clinical standpoint.

Over the past decade, Illumina short-read sequencing platforms have become a popular technology to elucidate the genomic content and molecular epidemiology of bacteria. However, the frequent occurrence of repeat elements prohibits the assembly of complete replicons (plasmids and chromosomes) and often results in hundreds of contigs per genome with an unclear origin. Plasmid and chromosome contigs are mingled in draft genome assemblies, which challenges the accurate reconstruction of plasmids. More recently, long-read sequencing platforms (Oxford Nanopore and PacBio) have successfully resolved this issue, but short-read sequencing remains the de facto standard in many microbiology laboratories [11,12,13,14].

Several fully automated bioinformatics tools are currently available to predict bacterial plasmids from short-read sequencing data. Since 2018, at least 15 different tools have been created for this purpose (Table S1). They can be broadly categorized into two main classes. The first class comprises software that produces a binary classification of contigs as either plasmid- or chromosome-derived, generating an output that predicts the complete plasmid content of a bacterial strain, often referred to as the ‘plasmidome’. An accurate plasmidome prediction has proven helpful to discover the genomic location of clinically relevant genes [15,16,17,18] and their role in shaping niche specificity [19], among others. The second class consists of tools that aim to recover distinct closed plasmid sequences. The output of these tools provides, in theory, a more comprehensive picture of the plasmid content of bacteria and allow to study the dissemination and epidemiology of specific plasmids [20].

Here, we reviewed the different tools and strategies to achieve binary prediction, for example fast k-mer based searches against reference plasmid databases (PlaScope and PlasmidSeeker), exploitation of the natural distribution bias of protein-coding genes between plasmids and chromosomes (Platon), and machine learning algorithms with different underlying features (cBAR, PlasFlow, ml-plasmids, PlasClass, RFPlasmid and PPR-Meta) and others. Furthermore, we benchmarked six tools aimed at reconstructing fully closed distinct plasmids for use with *E. coli*, by using complete *E. coli* genomes that were recently deposited to public databases. The strategies applied by the reconstruction tools consist of graph-based approaches (plasmidSPAdes, gplas), reference-based approaches (MOB-Suite, FishingForPlasmids) and hybrid approaches which use reference- and graph information (HyAsP and SCAPP). We assessed their performance based on their ability to correctly recover different plasmids as distinct and complete predictions, including plasmids that carry clinically relevant antibiotic resistance determinants, such as extended-spectrum beta-lactamase (ESBL) genes.

## Materials and Methods

### Review of plasmid prediction tools

We performed a systematic search of peer-reviewed publications deposited in PubMed by August 25th 2020, using the following search terms:

```
((plasmid*[Title])) AND ((software[Title/Abstract] ) OR (tool*[Title/Abstract]) OR program[Title/Abstract])) AND ((predict*[Title/Abstract] ) OR (sequencing[Title/Abstract] ) OR (identification[Title/Abstract] ) OR (prediction[Title/Abstract] ) OR (contigs[Title/Abstract] ) OR (assembly [Title/Abstract] ) OR (NGS[Title/Abstract])).
```

This search resulted in 238 peer-reviewed publications that we manually curated to obtain a list of 17 different tools with the goal to study the plasmid content of bacteria *in silico* (Table S1).

In order to find tools deposited on GitHub and GitLab, we used the search term “\*plasmid\*”. This resulted in 229 repositories from which 7 relevant tools were added to the selection (Table S1). The Github location of FishingForPlasmids was obtained through personal communication with the developer.

### Retrieving *E. coli* complete genomes and metadata from NCBI database

Ncbi-genome-download v0.2.10 (<https://github.com/kblin/ncbi-genome-download/>) was used to download all *E. coli* sequence labeled as ‘complete genomes’ up to August, 25th 2020 (n=1755). Metadata of the isolates was retrieved and parsed using Entrez-utilities v13.9 [21]. All scripts used to carry out the analyses in this study are available in a Git repository ([https://gitlab.com/jpaganini/recovering\\_ecoli\\_plasmids](https://gitlab.com/jpaganini/recovering_ecoli_plasmids)).

### Phylogenetic analysis

Phylogroups were determined *in silico* by using ClermonTyping v1.4.0 [22]. Core- and accessory-genome distances were calculated by using PopPUNK v1.2 [23] with standard parameters. PopPUNK was also used to build a core-genome neighbor-joining tree with 1381 complete *E. coli* genomes downloaded from the NCBI database on August, 25th 2020. Tree visualization and metadata information were integrated in Microreact [24] (Table S2).

### Benchmark data set selection

Isolates that were not sequenced by both long- and short-read technologies (n=559) were excluded, as well as sequences that were predicted as *Escherichia* cryptic clades [25] by *in silico* ClermonTyping (n=12) and genomes that exhibited a predicted accessory-genome distance larger than 0.5 by PopPUNK (n=2). We used a script written in R (version= 3.6.1) to remove genomes that had been used for developing the tested tools (n=601). Moreover, we excluded genomes that did not carry any plasmids (n=170), except for 19 randomly selected *E. coli* isolates without plasmids that were included as negative controls. In order to get a balanced data set, we removed a random sample of genomes isolated from farm animals (n=161). Finally, we removed 30 genomes containing short-read-only assembled contigs that did not align to any replicon in their respective closed reference genome. The data set resulted in 240 *E. coli* complete genomes, which carried a total of 631 plasmids (Figure S1, Table S3).

### Evaluating plasmid diversity in benchmarking data

We used Mash v2.2.2 ( $k=21$ ,  $s=1000$ ) to estimate the pairwise  $k$ -mer distances of all plasmids ( $n=3,264$ ) from all complete *E. coli* genomes ( $n=1,381$ ). The obtained distances were clustered using the  $t$ -distributed stochastic neighbour embedding ( $t$ -SNE) algorithm with a perplexity value of 30, and data points (which represents individual plasmid sequences) were coloured in orange if they were part of the benchmarking data set.

### Plasmid Predictions

Illumina raw reads were downloaded using SRA Tools (v2.10.9). Reads were trimmed using trim-galore (v0.6.6) (<https://github.com/FelixKrueger/TrimGalore>) to remove adapter contamination and bases with a phred quality score below 20. SPAdes (v3.14.0) [26] was applied to perform *de novo* assembly in careful mode and using  $k$ mer lengths of 37, 57 and 77. For isolates GCA\_014117345.1\_ASM1411734v1, GCA\_006352265.1\_ASM635226v1 and GCA\_003812945.1\_ASM381294v1, SPAdes was run using the `--isolate` option. The resulting contigs, assembly graphs and trimmed-reads were used as input for the different plasmid reconstruction tools, following the input requirements of the respective tools (Table S1). All tools were run with default parameters. Tool's versions were: FishingForPlasmids (no version information), MOB-suite (v3.0.0), SCAPP (v0.1.3), plasmidSPAdes (v3.14.0), gpLas (v0.6.1), HyAsP (v1.0.0).

### Analysis of the plasmid bins composition

We used QUAST (v5.0.2) to align the contigs of each bin to the respective closed reference genome. An extended description of the parameters used is available at Supplementary Material. Based on the alignment results, we calculated precision, recall and F1-score as specified below.

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

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

$$\text{F1-Score (bp)} = \frac{2 \times \text{Precision (bp)} \times \text{Recall (bp)}}{\text{Precision (bp)} + \text{Recall (bp)}}$$

If a bin was composed of contigs derived from different plasmids, precision, recall and F1-score were reported for each plasmid-bin combination.

In order to quantify the chromosomal sequence content (if any) on a bin, we defined a chromosome contamination metric as follows.

$$\text{Chromosome contamination} = \frac{\text{Alignment length against chromosome (bp)}}{\text{Total length of predicted bin (bp)}}$$

### Evaluating maximum theoretical recall for each reference plasmid

Depending on the input requirement of the respective tools (graph or contigs), we converted assembly graph nodes to FASTA format using the tool Any2Fasta (<https://github.com/tseemann/any2fasta>) or used the contigs produced by SPAdes and aligned them to their respective closed reference genomes using QUAST. Based on these alignments we calculated the maximum recall that could be obtained for reconstruction of every reference plasmid using short-read sequencing data (Supplementary Material).

### Antibiotic resistance gene (ARG) prediction

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

### Evaluating reconstruction of ARG plasmids

For bins that carried ARGs, we calculated RecallARG, as indicated below.

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

Bins that included the complete ARG content of the reference plasmid (RecallARG=1) and were linked to the correct plasmid backbone (F1-score>=0.95) were considered as correct reconstructions of the ARG-plasmid.

## Results

### Computational methods to predict the plasmidome or distinct plasmids

We used a systematic search of peer-reviewed publications and two popular software-repository hosting web services and retrieved a total of 25 plasmid- or plasmidome- prediction tools (Table S1). Most of the tools (n = 24) were fully automated and harbored the potential to be included in computational pipelines. Of these 24 tools, 13 tools were designed to analyze the plasmidome of multiple species using whole-genome sequencing data as input, while 8 tools can be applied to metagenomic sequences. A total of two tools, Recycler and RFPlasmid, worked with both types of input. Notably, we found one tool (FishingForPlasmids) that was developed to exclusively study the plasmid content of *E. coli*.

Based on the output, most of the tools (n = 23) can be broadly categorized into one of the following three classes. The first class comprises software that predicts the plasmidome, thus producing a binary classification of contigs as either plasmid- or chromosome-derived (n = 10). The second class consists of tools that aim to recover distinct plasmid sequences (n = 11) (Figure 1, Table S1). The third class of tools seeks to facilitate the detection of known plasmids (n = 2). Below, we briefly review the computational strategies applied by 17 tools that belong to the first two categories. Four tools were excluded from this review for distinct reasons: plasmIDent uses long-reads as input, plasmidID and plasmidAssembler use a similar approach to MOB-suite for plasmid reconstruction and PLACNET requires manual intervention from the user.



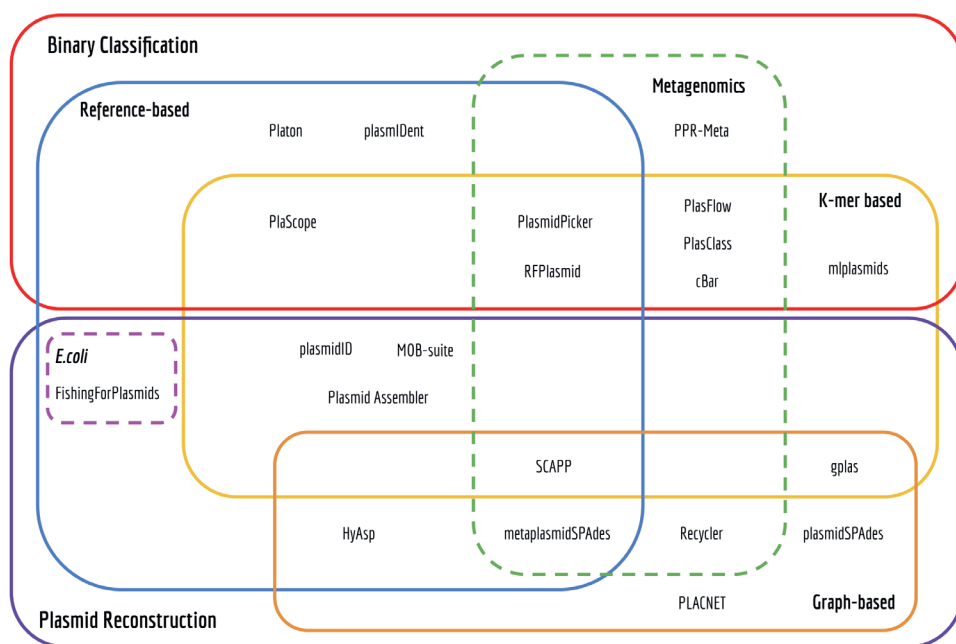


Figure 1. Euler diagram of bioinformatics tools to predict the plasmidome of bacteria.

### Binary Classification Tools

Binary classification tools take previously assembled contigs as input and classify them as being plasmid- or chromosome-derived.

PlaScope [27] and PlasmidPicker perform k-mer searches against reference plasmid databases. This strategy is very fast but limited to detecting k-mers that are present in the underlying database. Consequently, this produced high specificity and precision values but lower recall in a study that included a benchmark of PlaScope [27,28].

CBar, PlasFlow and PlasClass all share a common underlying principle: using short k-mer frequencies and machine learning (ML) algorithms to classify metagenomic assemblies. More specifically, cBAR relies on observed differences in pentamer frequencies and uses a sequential minimal optimization (SMO) model. PlasFlow calculates the frequencies of multiple k-mers sizes (between 5 and 7 nt) and utilizes a neural-network voting classifier to integrate predictions. PlasFlow has a better performance than cBAR [29,30], but shows less reliable results for short contigs [31]. PlasClass addresses this issue by using a set of four logistic regression classifiers, each trained on sequences of different length [31]. Similar to cBAR, mlplasmids also relies on pentamer frequencies but uses a Support Vector Machine (SVM) model to determine the origin of contigs for a single species, and contains models for *Escherichia coli*, *Klebsiella pneumoniae* and *Enterococcus faecium*. Mlplasmids outperformed both cBAR and PlasFlow when classifying data derived from whole-genome sequencing experiments, and it can also accurately predict the plasmid localization of several antimicrobial resistance genes [29]. RFPPlasmid [32], a recently released tool, uses a random forest classifier trained

with a hybrid approach by identifying chromosomal and plasmids marker genes using two databases and also pentamer frequencies. This tool also works with metagenomic assemblies, albeit only for contigs from the 17 different species for which classifiers were trained. Platon exploits the natural distribution bias of protein-coding genes between plasmids and chromosomes and also analyzes higher-level characteristics of the contigs: circularization, presence of replication and mobilization proteins, presence of oriT and incompatibility sequences [28].

Finally, PPR-Meta [33] allows simultaneous identification of both phages and plasmids fragments from metagenomes by using a Convolutional Neural Network. Notably, instead of k-mer frequencies, this tool uses one-hot matrices to represent nucleotides and amino-acids sequences [33].

Despite the differences in approaches and performances, none of the aforementioned tools attempted to further sort the predicted plasmidome into individual plasmids. As a consequence, these tools are not suitable for studying the epidemiology of specific plasmids.

### Plasmid reconstruction tools

Based on their computational strategies, we can roughly subdivide plasmid reconstruction tools into three different categories: (i) *de novo* reconstruction of plasmids using assembly graph information, (ii) reference-based approaches and (iii) hybrid approaches.

PlasmidSPAdes, Recycler, metaplasmidSPAdes and gplas [34,35,36] perform a *de novo* reconstruction of plasmids using assembly graph information. PlasmidSPAdes and Recycler were released in 2016 and were the first tools that exploited the information on the assembly graph for identifying individual plasmids. PlasmidSPAdes is based on the assumption that plasmids have a different copy number than the chromosome, and therefore plasmid contigs will exhibit a different read coverage than chromosomal contigs. A number of studies have shown that this tool is able to reconstruct bacterial plasmids with high recall [11,37,38], but they have also revealed two major disadvantages of this approach: (1) plasmidSPAdes fails to identify large plasmids that have the same copy number as the chromosome and (2) it has a tendency to merge different plasmids together. Recycler also tries to identify plasmid-paths in the assembly graph by using coverage information but incorporates additional data regarding the topology of the selected paths. The main rationale behind this algorithm is that selected plasmid-paths should be cyclic, coverage should be homogeneous amongst all contigs and mated pair-end reads should map to the same path. Recycler appears to successfully identify short plasmids but yields very low precision values for long plasmids [11,37]. This issue is partially addressed by metaplasmidSPAdes, released in 2019 as an improvement on the original prediction algorithm of plasmidSPAdes. This tool allows prediction of dominant plasmids in metagenomes, defined as plasmids with coverage exceeding that of chromosomes and other plasmids. The algorithm iteratively extracts cyclic subgraphs with increasing coverage from the metagenome assembly graph. These potential plasmid sequences are later analyzed by a naive Bayesian classifier, called plasmidVerify, that further assesses the gene content of potential plasmids. None of the aforementioned tools takes advantage of the information embedded in the nucleotide sequences of the assembled contigs to a priori simplify the task of identifying plasmid subgraphs. In contrast, gplas initially classifies assembled contigs as plasmid-derived or chromosome-derived by using mlplasmids (or plasflow), a tool that exploits short k-mer frequencies for achieving such classification. Subsequently, plasmid-derived unitigs act as seeds for finding plasmid-walks with homogeneous coverage in the assembly graph, using a greedy

approach. Gplas generates a plasmidome network in which nodes corresponding to plasmid units and edges are created and weighted based on the co-existence of the nodes in the solution space of the computed walks. Finally, this plasmidome network is queried by a selection of network partitioning algorithms for generating bins of contigs that belong to the same plasmid [36].

MOB-suite and FishingForPlasmids use a reference-based approach for reconstructing individual plasmids. MOB-suite works as a modular set of tools for clustering, reconstruction and typing of plasmids from assemblies. This software initially uses Mash [39] and a single-linkage clustering algorithm to create clusters of similar plasmids present in a reference database. Input contigs are then aligned against this database using Blast and assigned to a plasmid cluster according to the best hits obtained. Contigs assigned to the same reference cluster constitute potential individual plasmid units. Also, the topology of the contigs is evaluated and every circular contig is considered an individual plasmid. Finally, each identified plasmid is queried against a different database for finding known replication and mobilization proteins and oriT sequences. According to the authors, MOB-suite performs better than plasmidSPades at correctly reconstructing plasmids from a benchmarking data set that included more than 370 plasmids from 14 different bacterial species [38]. However, the authors identified that MOB-suite splits single plasmids into different predictions more often than plasmidSPades. FishingForPlasmids attempts to reconstruct individual plasmids from *Escherichia coli* assemblies. This tool identifies plasmid-contigs by using BlastN to align each contig against a curated *E. coli* database. Each plasmid-derived sequence is further classified into discrete components by using a combination of plasmidFinder and pMLST [14].

Finally, HyAsP and SCAPP use a hybrid approach, mixing principles from reference-based and *de novo* methods. In HyAsP, a set of potential plasmid contigs is first selected based on: (1) a high density of known plasmid genes, identified by using a database, (2) high read coverage and (3) a length that does not exceed a maximum threshold. These plasmid-contigs will be used as seeds for finding plasmid-walks within the original assembly graph using a greedy algorithm. Plasmid-walks must satisfy the following conditions: (1) have a uniform GC content and sufficient read coverage, (2) do not have large gene-free segments and (3) total length of the plasmid-walk does not exceed a threshold. SCAPP, on the other hand, is designed for finding plasmids in metagenome assemblies. This algorithm starts by finding potential plasmid-contigs based on two strategies: (1) searching for plasmid-specific genes by using a curated database and (2) assigning weight to each contig based on the output from PlasClass, a ML-based binary classifier. The assembly graph is then queried to find cyclic walks of uniform coverage, similar to Recycler, but prioritizing the inclusion of contigs with strong evidence of plasmid-origin [40].

### The benchmark data set represents the diversity of sequenced plasmids

To benchmark the aforementioned plasmid reconstruction tools, we used a data set of 240 *E. coli* strains with complete genome sequences and short read data available from public databases that harbored 631 plasmids. These *E. coli* genomes were absent from all training data sets used to develop the selected plasmid prediction tools. The majority of the genomes derived from Europe ( $n = 170$ ), Asia ( $n = 39$ ) and North America ( $n = 24$ ) (Figure 2A). They were isolated from multiple sources such as animals ( $n = 103$ ), humans-clinical samples ( $n = 27$ ), humans-community samples ( $n = 4$ ), environmental sources ( $n = 86$ ) and unknown sources ( $n = 13$ ) (Figure 2B).

To assess if the selected genomes were a representative sample of the phylogenetic diversity of *E. coli*, we built a neighbor-joining tree combining our data set with 1141 complete *E. coli* genomes and determined the phylogroup of each of these genomes *in silico*. This analysis revealed that the selected genomes were distributed across the core-genome tree and that all phylogroups were represented with at least five strains. (Figure 2C).

Most of the genomes carried one ( $n = 73$ ), two ( $n = 49$ ) or three ( $n = 28$ ) plasmids, but notably some genomes contained as much as nine ( $n = 3$ ), ten ( $n = 1$ ) or eleven ( $n = 1$ ), with a median of two (mean = 2.62 plasmids). We found a clear bimodal plasmid size distribution, with peaks around 4500 bp and 100,000 bp (Figure 2D). Consequently, plasmids with a length smaller than 18,000 bp were classified as ‘small’ ( $n = 273$ ), while plasmids that exceeded this cut-off value were classified as ‘large’ ( $n = 358$ ).

Next, we wanted to assess the diversity of plasmids included in the benchmark data set. We used Mash to estimate the pairwise k-mer distances of all plasmids ( $n = 3264$ ) from all complete *E. coli* genomes ( $n = 1381$ ) and clustered them with the t-SNE algorithm. Plasmids included in this study were distributed among all major clusters, suggesting that this data set is able to properly capture the diversity of the *E. coli* pan-plasmidome currently available at NCBI (Figure 2E).

### **A third of all plasmids could not be correctly reconstructed by any of the tools**

We selected six tools to reconstruct distinct plasmid sequences. These tools applied different computational strategies: graph-based (plasmidSPAdes, gplas), reference-based (MOB-Suite, FishingForPlasmids) and hybrid (HyAsP and SCAPP).

The rest of the plasmid reconstruction tools were not included in the analysis because of a variety of reasons: Plasmid Assembler couldn’t be installed, plasmidID predictions were not completed due to errors during execution, PLACNET required manual intervention of the user, Recycler provided suboptimal results in comparison with plasmidSPAdes and HyAsP in previous studies [11,37] and metaplasmidSPAdes uses a similar approach to plasmidSPAdes but optimized for metagenomic samples.

We evaluated the predictions obtained with the six selected plasmid reconstruction tools in terms of (i) speed and memory requirements, (ii) the number of plasmid predictions, (iii) correct reconstruction of reference plasmids, (iv) chromosomal contamination included in predicted plasmids, and (v) correct reconstruction of ARG-plasmids.

We used a High-Performance Cluster (HPC) to run the tools with minimal resources (number of cores = 2, 4GB of RAM per genome), and documented the total CPU-time and memory required by each of them (Table 1, Figure S2). Most tools required less than 100 CPU hours to complete all predictions, except for plasmidSPAdes which used 321.07 CPU hours. In contrast, FishingForPlasmids was the fastest tool and completed the task in 10.60 CPU hours. PlasmidSPAdes and SCAPP had the highest memory requirements, utilizing a total of 442.03 Gb and 435.23 Gb of RAM, respectively. The remaining tools required less than 300 Gb to complete all predictions. Notably, FishingForPlasmids only required a total of 36.57 Gb.



**Figure 2.** (A) Genomes distribution according to geographical location and (B) isolation source. (C) Core-genome clustering constructed using PopPUNK. We included 1381 complete *E. coli* genomes available at NCBI database. Orange tips ( $n = 240$ ) indicate genomes that were included in the benchmarking data set, and outer colors indicate phylogroups. (D) Plasmid length histogram and density plot. Dashed line indicates the cut-off length (18,000 bp) for considering a plasmid as small or large (E) t-SNE plot based on plasmids k-mer distances obtained with MASH ( $k = 21$ ,  $s = 1000$ ). Plasmids included in this benchmark ( $n = 631$ ) are colored in orange.

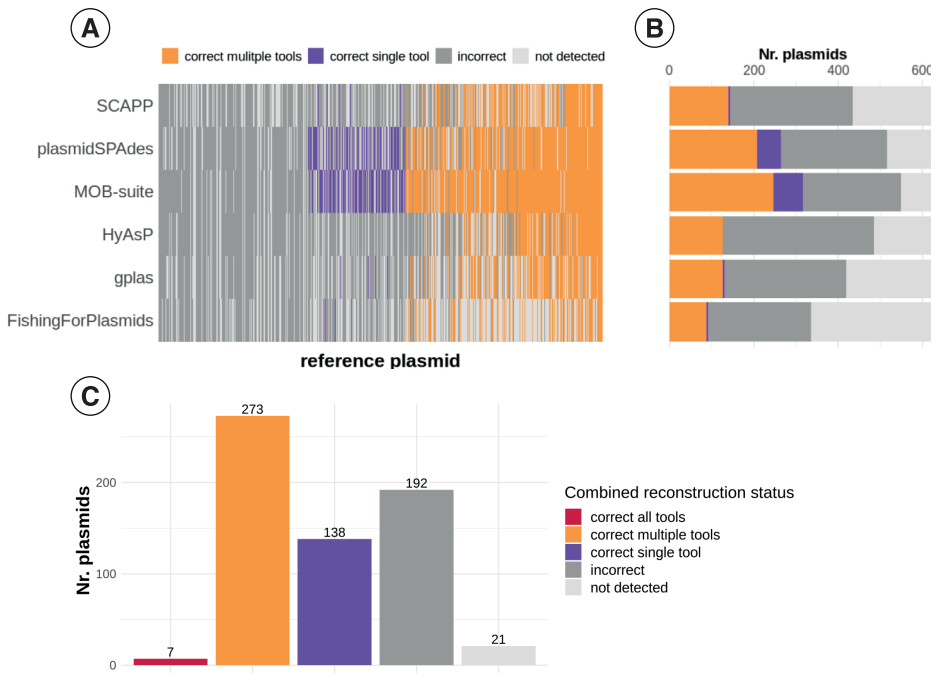
Next, we evaluated the number of plasmid predictions produced by each tool and calculated the difference between this number and the true number of plasmids present in the benchmark data set (Table 1, Figure S3). The total number of plasmid predictions ranged from 377 (FishingForPlasmids) to 2590 (HyAsP). PlasmidSPAdes, MOB-suite, SCAPP and HyAsP overestimated the true number of plasmids ( $n = 631$ ), while gplas and FishingForPlasmids underestimated this number. PlasmidSPAdes displayed the least deviation by producing 642 bins, and therefore exceeding the total number of plasmids by 11. Nevertheless, these absolute numbers do not reflect whether predictions were correct or incorrect.

In order to evaluate how the different tools performed at recovering *E. coli* plasmids as distinct and complete predictions, we studied the distributions of recall, precision and F1-score (Table 1, Figure S4A–C) for all plasmid predictions made by the tools. Based on these results, we determined an F1-score cut-off value of 0.95 to define a plasmid as correctly reconstructed (or recovered) (Figure S4D).

MOB-suite correctly recovered 317 (50.24%) plasmids (F1-score  $\geq 0.95$ ), including 70 (11.10%) that couldn't be reconstructed by any other software (Figure 3A,B, Table 1). Similarly, plasmidSPAdes reconstructed a total of 263 (41.68%) plasmids, including 55 (8.72%) that were not recovered by other tools. Interestingly, 14 of these 'unique reconstructions' were also missing from the short-read assembly graphs (Supplementary Materials, Tables S4 and S5). The rest of the tools achieved smaller quantities of correct plasmid reconstructions, with values ranging from 92 (14.58%) to 152 (24.09%) (Figure 3A,B, Table 1).

We found that a total of 418 (66.25%) plasmids were correctly reconstructed by at least one of the tools (Figure 3C). Out of these, only 7 (1.11%) were reconstructed by all tools concurrently, 273 (43.26%) by multiple tools and 138 (21.9%) by a single tool. Interestingly, combining MOB-suite and plasmidSPAdes predictions together achieved the correct reconstruction of 400 (63.39%) plasmids, and incorporating the predictions from the remaining tools only resulted in the reconstruction of 18 (2.85%) additional plasmids. Notably, a total of 213 (33.75%) plasmids were incorrectly reconstructed (F1 score  $< 0.95$ ) by all tools, including 21 (3.32%) that were not even detected. The majority of ARG-plasmids ( $n = 85$ , 57.8%) could not be correctly reconstructed by any of the tools (Table S6).

We also compared the performance of the software when attempting to reconstruct small- and large plasmids separately. For small plasmids, we discovered that all tools displayed similar F1-score distributions, with medians ranging from 0.95 to 0.99. However, the tools did not detect 21.25–89.74% of small plasmids (Figure S6A,B). PlasmidSPAdes and MOB-suite were the only tools that achieved the correct reconstruction of most of these replicons, with a total of 168 (61.54%) and 155 (55.31%), respectively (Table 1). When considering the reconstruction of large plasmids, percentages of not-detected plasmids were much lower and ranged from 2.23% to 20.11% across tools. MOB-suite exhibited the highest F1-score values (median = 0.74, IQR = 0.17–0.97) and correctly reconstructed 166 (46.3%) of these replicons, significantly surpassing the reconstruction capacity of the rest of the tools, which ranged from 45 (12.57%) to 95 (26.54%) (Table 1, Figure S6A,B). Not surprisingly, most tools correctly reconstructed a higher fraction of small plasmids, and also displayed higher F1-score values (Table 1, Figure S6A,B) when comparing with the reconstruction of large plasmids. FishingForPlasmids was the only exception as it recovered a total of 14 (5.13%) small and 78 (21.79%) large plasmids.



**Figure 3.** (A) Reconstruction performance of each tool for all reference plasmids. Reference plasmids have been ordered according to the number of tools by which they were correctly reconstructed, from low (left; reconstructed by 0/6 tools) to high (right; reconstructed by 6/6 tools). Plasmids that were reconstructed with an F1-score  $\geq 0.95$ , were considered as correct reconstructions. Plasmids for which no contig was included in the predictions were considered as ‘not-detected’. (B) Absolute count of all reconstruction status achieved by each tool. (C) Absolute count of reconstruction categories when combining predictions from all tools.

All tools incorrectly incorporated chromosome-derived sequences into their predictions (Figure S7, Table 1). FishingForPlasmids performed best at avoiding this error, and only 7 (1.8%) predictions contained chromosomal contamination. In contrast, HyAsP introduced chromosomal contigs in 1340 (51.7%) predictions with a chromosome contamination median of 0.88 (IQR = 0.5–0.99), including 1251 pure chromosome bins (chromosome contamination = 1). Notably, plasmidSPAdes and MOB-suite had a similar proportion of contaminated bins, 295 (46%) and 297 (40.2%), yet with different chromosome contamination medians of 0.75 (IQR = 0.14–0.92) and 0.10 (IQR = 0.03–0.99), respectively. Out of these, MOB-suite produced 65 predicted bins which exclusively consisted of chromosome sequences, while plasmidSPAdes generated 20 of them. SCAPP introduced chromosomal sequences in 249 (25.2%) predictions, but notably only 1 of them was only composed of chromosome sequences. Finally, gplas incorporated chromosomal sequences in 197 (35.8%) predictions, of which 70 were exclusively composed of these types of sequences.

## Chapter 2

**Table 1.** Summary of tools performances.

	HyAsP	MOB-suite	gplas	plasmidSPAdes	SCAPP	FishingFor Plasmids
<b>Computational Performance</b>						
Memory Usage (GB)	299.2	202.82	150.36	442.03	435.23	36.57
CPU-Time (hr)	46.57	46.62	83.64	321.07	70.96	10.6
<b>Nr. of plasmid predictions</b>						
Nr. total predicted plasmids (bins)	2590	738	550	642	986	377
Nr. correct predictions of plasmid absence (%)	2 (10.53)	13 (68.42)	17 (89.47)	9 (47.37)	17 (89.47)	18 (94.74)
<b>Plasmids reconstruction All Plasmid (n = 631)</b>						
<u>Nr. correctly reconstructed plasmids (%)</u>	127 (20.13)	317 (50.24)	130 (20.6)	263 (41.68)	152 (24.09)	92 (14.58)
Nr. small plasmids (%)	82 (30.04)	151 (55.31)	87 (31.87)	168 (61.54)	98 (35.9)	14 (5.13)
Nr. large plasmids (%)	45 (12.57)	166 (46.37)	43 (12.01)	95 (26.54)	54 (15.08)	78 (21.79)
<u>Nr. incorrectly reconstructed plasmids (%)</u>	358 (56.74)	231 (36.61)	289 (45.8)	252 (39.94)	291 (46.12)	243 (38.51)
Nr. small plasmids (%)	53 (19.41)	50 (18.32)	17 (6.23)	47 (17.22)	59 (21.61)	14 (5.13)
Nr. large plasmids (%)	305 (85.20)	181 (50.56)	272 (75.98)	205 (57.26)	232 (64.80)	229 (63.97)
<u>Nr. undetected plasmids (%)</u>	146 (23.14)	83 (13.15)	212 (33.6)	116 (18.38)	188 (29.79)	296 (46.91)
Nr. small plasmids (%)	138 (50.55)	72 (26.37)	169 (61.9)	58 (21.25)	116 (42.49)	245 (89.74)
Nr. large plasmids (%)	8 (2.23)	11 (3.07)	43 (12.01)	58 (16.2)	72 (20.11)	51 (14.25)
<u>F1-score (median - IQR)*</u>	0.12 (0.04 - 0.41)	0.89 (0.3 - 0.98)	0.59 (0.3 - 0.94)	0.95 (0.49 - 0.99)	0.18 (0.07 - 0.81)	0.64 (0.29 - 0.93)
Small plasmids*	0.98 (0.76 - 0.99)	0.98 (0.94 - 0.99)	0.99 (0.98 - 0.99)	0.98 (0.96 - 0.99)	0.96 (0.88 - 0.99)	0.95 (0.7 - 0.98)
Large plasmids*	0.11 (0.04 - 0.32)	0.74 (0.17 - 0.97)	0.49 (0.21 - 0.76)	0.6 (0.31 - 0.97)	0.12 (0.06 - 0.41)	0.61 (0.28 - 0.91)
<u>Recall (median - IQR)*</u>	0.07 (0.02 - 0.32)	0.89 (0.21 - 0.99)	0.5 (0.22 - 0.93)	0.99 (0.88 - 1)	0.13 (0.04 - 0.78)	0.51 (0.18 - 0.93)
Small plasmids*	1 (0.92 - 1)	1 (0.96 - 1)	1 (0.98 - 1)	1 (1 - 1)	0.99 (0.92 - 1)	1 (0.96 - 1)
Large plasmids*	0.06 (0.02 - 0.2)	0.63 (0.12 - 0.96)	0.4 (0.16 - 0.72)	0.94 (0.36 - 0.99)	0.07 (0.03 - 0.31)	0.46 (0.16 - 0.84)
<u>Precision (median - IQR)*</u>	0.87 (0.5 - 0.98)	0.98 (0.68 - 1)	0.97 (0.55 - 1)	0.93 (0.41 - 0.98)	0.8 (0.39 - 0.94)	1 (1 - 1)
Small plasmids*	0.96 (0.86 - 0.98)	0.98 (0.95 - 0.99)	0.98 (0.97 - 0.99)	0.96 (0.92 - 0.98)	0.95 (0.83 - 0.98)	0.96 (0.65 - 0.97)
Large plasmids*	0.84 (0.48 - 0.98)	0.97 (0.53 - 1)	0.93 (0.47 - 1)	0.58 (0.33 - 0.99)	0.75 (0.34 - 0.92)	1 (1 - 1)
<u>Chromosome contamination (Median - IQR)</u>	0.88 (0.59 - 0.99)	0.1 (0.03 - 0.99)	0.45 (0.11 - 1)	0.75 (0.14 - 0.92)	0.3 (0.09 - 0.66)	1 (0.6 - 1)
Nr. bins with chromosome contamination (%)	1,340 (51.73)	297 (40.2)	197 (35.81)	295 (45.95)	249 (25.25)	7 (1.86)



Table 1. Cont.

	HyAsP	MOB-suite	gplas	plasmidSPAdes	SCAPP	FishingFor Plasmids
Nr. pure chromosome bins	1251	65	70	20	1	4
<b>Plasmids reconstruction ARG-plasmids (n=147)</b>						
<u>ARGs in bins</u>						
Nr. plasmid-derived ARGs (%)	525 (84.95)	548 (88.67)	331 (53.56)	390 (63.11)	223 (36.08)	133 (21.52)
Nr. chromosome-derived ARGs	130	92	71	29	34	1
<u>Reconstruction status</u>						
Nr. plasmids correctly reconstructed (%)	5 (3.4)	41 (27.89)	10 (6.8)	23 (15.65)	10 (6.8)	13 (8.84)
Nr. plasmids predicted with incorrect backbones (%)	62 (42.18)	49 (33.33)	38 (25.85)	59 (40.14)	23 (15.65)	9 (6.12)
Nr. plasmids predicted with incomplete ARG content (%)	66 (44.9)	47 (31.97)	59 (40.14)	28 (19.05)	39 (26.53)	28 (19.05)
Nr. plasmids with no ARGs predicted (%)	14 (9.52)	10 (6.8)	40 (27.21)	37 (25.17)	75 (51.02)	97 (65.99)
<u>Large ARG-plasmids reconstruction metrics (n=143)</u>						
Recall (Median - IQR)*	0.06 (0.02 - 0.16)	0.38 (0.09 - 0.88)	0.29 (0.14 - 0.62)	0.87 (0.2 - 0.96)	0.06 (0.03 - 0.17)	0.35 (0.15 - 0.55)
Precision (Median - IQR)*	0.84 (0.46 - 0.99)	0.92 (0.42 - 1)	0.86 (0.44 - 1)	0.47 (0.31 - 0.92)	0.71 (0.32 - 0.88)	1 (1 - 1)
F1-score (Median - IQR)*	0.1 (0.04 - 0.26)	0.44 (0.13 - 0.9)	0.41 (0.19 - 0.65)	0.53 (0.24 - 0.73)	0.1 (0.05 - 0.26)	0.51 (0.25 - 0.69)
Nr. detected plasmids (%)	141 (98.60)	141 (98.60)	135 (94.41)	129 (90.21)	113 (79.02)	138 (96.50)
<b>Plasmids reconstruction ESBL-plasmids (n = 60)</b>						
<u>ESBL genes in bins</u>						
Nr. plasmid-derived (%)	52 (86.67)	57 (95)	27 (45)	40 (66.67)	23 (38.33)	11 (18.33)
Nr. chromosome-derived (%)	10	8	7	2	2	0
<u>Reconstruction status</u>						
Nr. ESBL genes in correct plasmid backbone (%)	0 (0)	20 (33.33)	4 (6.67)	10 (16.67)	5 (8.33)	6 (10)
Nr. ESBL genes in incorrect plasmid backbone (%)	52 (86.67)	37 (61.67)	23 (38.33)	30 (50)	18 (30)	5 (8.33)
F1-score (Median - IQR)*	0.29 (0.07 - 0.46)	0.93 (0.72 - 0.97)	0.69 (0.45 - 0.88)	0.65 (0.51 - 0.95)	0.27 (0.09 - 0.84)	0.98 (0.71 - 0.99)
Recall (Median - IQR)*	0.18 (0.04 - 0.31)	0.89 (0.77 - 0.96)	0.65 (0.36 - 0.84)	0.96 (0.89 - 0.97)	0.23 (0.05 - 0.84)	0.95 (0.56 - 0.99)
Precision (Median - IQR)*	0.91 (0.54 - 0.98)	0.98 (0.93 - 1)	0.97 (0.89 - 1)	0.52 (0.38 - 0.95)	0.85 (0.72 - 0.92)	0.99 (1 - 1)

\* In all cases, undetected plasmids were not included in the calculation of Precision, Recall and F1-score.

### Plasmids carrying antibiotic resistance genes were difficult to reconstruct for all tools

Our data set included 147 (23.3%) plasmids containing antibiotic resistance genes (ARG-plasmids), carrying a total of 618 resistance genes. Most of these replicons carried one ( $n = 43$ ), two ( $n = 17$ ), three ( $n = 12$ ) or four ( $n = 17$ ) ARGs (Table S6). Interestingly, plasmids carrying ARGs had a median length of 109,773 bp (IQR = 83,300–132,865 bp), and were markedly larger than plasmids with no resistance determinants (median length 6930 bp; IQR = 4072–91,111 bp). Furthermore, 143 (97.2%) ARG-plasmids were classified as large, while only 4 (2.8%) were small plasmids (Figure S8A).

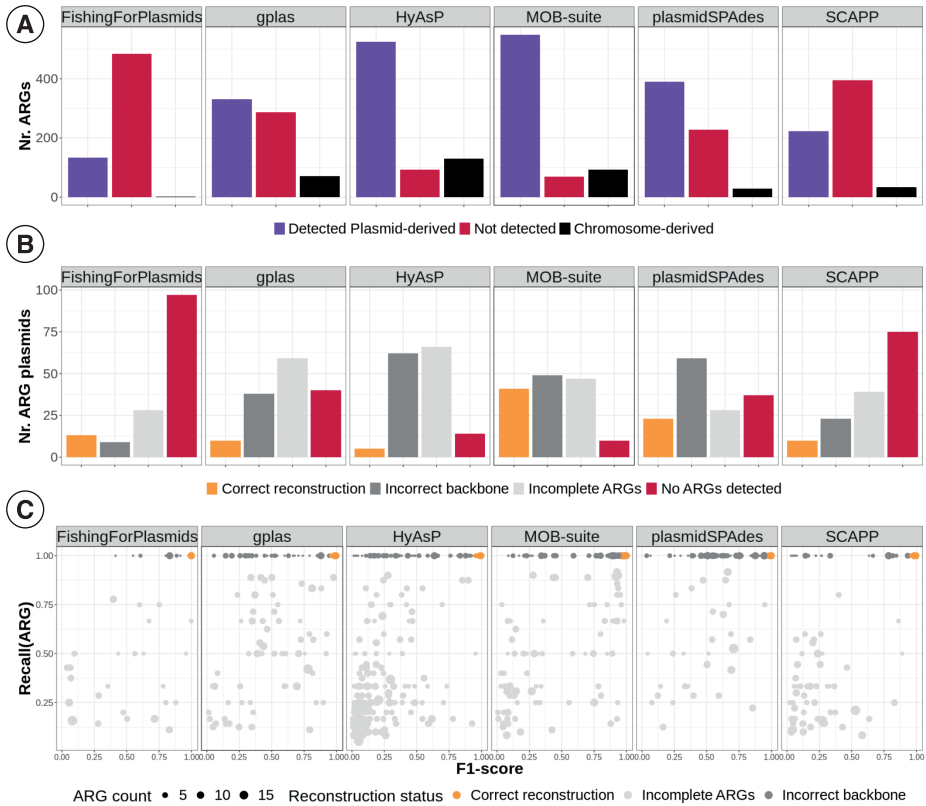
To investigate how the tools performed at reconstructing ARG-plasmids, we analyzed Recall, Precision and F1-score values for these replicons (Figure S8B–D). Furthermore, we extracted the bins that contained antibiotic resistance genes, and explored the fraction of detected ARGs in each prediction -Recall(ARG)-. An ARG-plasmid was considered as correctly reconstructed if the prediction simultaneously included all ARGs -Recall(ARG) = 1- and correctly represented the reference plasmid backbone (F1-score  $\geq 0.95$ ).

We discovered that the reconstruction of large ARG-plasmids was particularly challenging for the evaluated tools, since all of them exhibited lower F1-score values in comparison with the reconstruction of large non-ARG-plasmids (Figure S8B,E, Table 1). We excluded small plasmids from this comparison due to the low amount of small ARG-plasmids present in our data set.

MOB-suite correctly identified 548 (88.67%) plasmid-derived ARGs, and achieved 41 (27.89%) correct ARG-plasmid reconstructions (Figure 4A,B, Table 1). In 49 (33.3%) additional predictions, all ARGs were assigned into a single bin -Recall(ARG) = 1-, but the bin incorrectly represented the reference plasmid backbone (F1-score  $< 0.95$ ) (Figure 4C) by being incomplete, hybridized with sequences derived from other replicons, or both (Figure S9). Moreover, we discovered that MOB-suite incorrectly incorporated 92 chromosome-derived ARGs, distributing them among 39 bins. Finally, we found that when predicting large ARG-plasmids, this tool presented remarkably lower recall values (median = 0.38, IQR = 0.09–0.88) in comparison with reconstruction of large non-ARG-plasmids (median = 0.87, IQR = 0.19–0.98) (Figure S8C).

PlasmidSPAdes detected 390 (63.11%) plasmid-derived ARGs, and correctly reconstructed 23 (15.65%) ARG-plasmids. Additionally, in 59 (40.14%) predictions all ARGs were assigned to a single bin, but the plasmid backbone was most frequently contaminated with sequences from other replicons (Figure S9). Notably, this tool couldn't predict any of the ARGs present in 37 (25.17%) reference ARG-plasmids (Figure 4A,B, Table 1). Finally, for the reconstruction of large ARG-plasmids, plasmidSPAdes presented remarkably lower precision values (median = 0.47, IQR = 0.31–0.92) in comparison with reconstruction of large non-ARG-plasmids (median = 0.9, IQR = 0.35–1) (Figure S8D).

The rest of the tools successfully reconstructed smaller fractions of ARG-plasmids, ranging from 5 (3.4%) to 13 (8.84%). Interestingly, HyAsP detected a high fraction of plasmid-derived ARGs ( $n = 525$ ,  $n = 84.95\%$ ), but it only achieved the correct reconstruction of 5 (3.4%) ARG-plasmids. For most HyAsP predictions, all ARGs couldn't be assigned to a single bin ( $n = 66$ , 44.9%) or presented an incorrect plasmid backbone ( $n = 62$ , 42.18%). FishingForPlasmids detected the least amount of resistance genes ( $n = 133$ , 21.52%) and couldn't predict any of the ARGs present in 97 (66%) reference ARG-plasmids.



**Figure 4.** (A) Bar plot displaying the number of plasmid-derived ARGs that were detected/not detected by each of the tools. This plot also shows the number of chromosome derived ARGs included in the plasmid predictions. (B) Bar plot displaying the number of reference ARG-plasmids belonging to each different reconstruction category. Reconstruction categories were defined as follows. Correct reconstruction: all ARGs were predicted in the same bin ( $\text{Recall}(\text{ARG}) = 1$ ) and the backbone of the plasmid was correct ( $\text{F1-score} \geq 0.95$ ). Incorrect backbone: all ARGs were predicted in the same bin ( $\text{Recall}(\text{ARG}) = 1$ ) but the backbone of the plasmid was incorrect ( $\text{F1-score} < 0.95$ ). Incomplete ARGs: Not all ARGs were included in the same bin ( $\text{Recall}(\text{ARG}) < 1$ ). No ARGs detected: None of the ARGs derived from the reference plasmids were included in any bins created by the tool. (C) Scatter-plot showing relation between  $\text{Recall}(\text{ARG})$  and  $\text{F1-score}$  (bp) values for predictions that carry at least one ARG of plasmid origin. Dots are colored according to the same criteria as in B.

Next, we evaluated the performance of the tools when reconstructing plasmids that carry ESBL genes (ESBL plasmids). Our data set included 60 ESBL plasmids, each carrying a single ESBL gene. Most abundant ESBL variants were CTX-M15 ( $n = 16$ , 25%), CTX-M55 ( $n = 12$ , 20%) and CTX-M1 ( $n = 6$ , 10%) (Figure S10A). Furthermore, we observed that ESBL genes were harbored by plasmids with diverse sequences (Figure S10B).

MOB-suite correctly identified a total of 57 (95%) ESBL genes of plasmid origin, of which 20 were also assigned to the correct plasmid backbone ( $\text{F1-score} \geq 0.95$ ), resulting in a 33% correct reconstruction of the ESBL plasmids (Table 1, Figure S11A). Despite this, MOB-suite predictions achieved high  $\text{F1-scores}$  for reconstruction of ESBL plasmids (median = 0.93, IQR = 0.72–0.97) (Table 1, Figure S11B).

The rest of the tools reconstructed ESBL plasmids with less success, ranging from 0 (0%) to 10 (16.67%) total correct reconstructions (Table 1, Figure S11A). HyAsP detected a high fraction of plasmid-derived ESBL genes ( $n = 52$ , 86.67%), but did not achieve the correct reconstruction of any plasmids. PlasmidSPAdes detected the majority of plasmid-derived ESBL genes ( $n = 40$ , 66.66%), and these were included in bins that presented high recall (median = 0.97, IQR = 0.77–0.96) but low precision values (median = 0.52, IQR = 0.38–0.95) (Table 1, Figure S11C).

## Discussion

A tool that is able to correctly predict *E. coli* plasmids will assist in identifying clinically relevant plasmids [41,42,43,44] and improve our understanding of the complex dynamics of ARG dissemination across different ecological niches [45,46,47]. From the vast offer of software to predict plasmids from short-read data we selected six tools and benchmarked their performances when attempting to reconstruct individual *E. coli* plasmids, with a special focus on plasmids that carry ARGs.

A total of 418 (66.24%) plasmids were correctly reconstructed by at least one of the tools compared in this benchmark. Interestingly, 400 (63.39%) of these plasmids were recovered by combining the predictions from MOB-suite and plasmidSPAdes alone. Therefore, adding the predictions from the rest of the tools resulted only in 18 (2.85%) additional correct reconstructions.

We observed that plasmidSPAdes correctly reconstructed the highest fraction of small plasmids ( $n = 168$ , 61.5%). This result is consistent with the observations that small plasmids usually have high copy numbers [48] and therefore exhibit a higher coverage; which in theory would facilitate their prediction using this tool. A similar success at predicting small plasmids was also reported by [11,38]. Nevertheless, it is worth noticing that most small plasmids ( $n = 215$ , 79%) are represented as a single node in the assembly graph. Therefore, using a binary classification tool would be sufficient for correctly predicting these replicons.

MOB-suite correctly reconstructed a total of 166 (46.37%) large plasmids, and considerably outperformed the rest of the tools, which ranged from 45 (12.57%) to 95 (26.54%) correct reconstructions. Nevertheless, MOB-suite's performance strongly depends on its underlying database, which is enriched for Enterobacteriaceae plasmid sequences [38]. Consequently, the reconstruction capacity of this tool could be different when attempting to predict plasmids from bacterial species less frequently represented in its database.

A third ( $n = 213$ , 33.76%) of all plasmids could not be correctly reconstructed by any of the evaluated tools. In particular, the reconstruction of ARG-plasmids proved to be problematic. We hypothesize that ARG-plasmids constitute a particularly hard puzzle to solve for all compared computational approaches, for several reasons.

Firstly, ARG-plasmids usually carry a high number of repeated sequences [49,50,51,52], and therefore exhibit highly entangled assembly graphs. Secondly, ARGs are frequently located on large plasmids with low copy number, and therefore have coverage values that are similar to chromosomes [48,52]. Consequently, finding plasmid-walks with differential coverage in the assembly graphs

could be challenging for all tools relying on this strategy. This hypothesis is supported by the observation that plasmidSPAdes predicted large ARG-plasmids with the lowest precision values (median = 0.47, IQR = 0.31–0.92) of all tools, indicating that these plasmids are more frequently merged with sequences derived from other replicons. Additionally, this tool failed to predict 37% of all plasmid-located ARGs, which would be explainable in case that these contigs should have coverage values similar to the chromosomes.

Thirdly, ARG-plasmids are frequently built as mosaic-like structures, containing mobile components that can be found in different plasmid backbones [48,52,53,54,55]. This type of genomic organization also complicates their reconstruction using reference-based methods, since databases might contain very similar fragments that are shared by a variety of plasmids. Consequently, unequivocally assigning these “shared fragments” to a unique reference plasmid (or plasmid group) could be problematic. This is supported by the results obtained using MOB-suite. This software identified the highest proportion of plasmid-derived ARGs ( $n = 548$ , 88.67%), but most ARG-plasmids reconstructions had either an incomplete ARG content ( $n = 47$ , 31.97%) or an incorrect backbone ( $n = 49$ , 33.33%). These results, in combination with the low recall values observed (median = 0.38, IQR = 0.09–0.88) seems to suggest that large ARG-plasmids were frequently split into multiple bins.

Despite the aforementioned limitations, MOB-suite was the most effective tool at predicting ARG-plasmids in *E. coli*, achieving the correct reconstruction of 41 (27.89%) of these, while the rest of the tools ranged from 5 (3.4%) to 23 (15.65%) correct ARG-plasmid reconstructions. Additionally, MOB-suite was the best performing tool for prediction of ESBL-plasmids. It identified 57 (95%) plasmid-borne ESBL-genes and had a median F1-score of 0.93 (IQR = 0.72–0.97). However, it must be noted that a fraction ( $n = 13$ , 22.80%) of ESBL-plasmid predictions presented low F1-score values, implying that in these cases the contigs carrying the ESBL gene were associated with the incorrect plasmid backbone.

All tools exhibited chromosomal contamination in their predictions. Notably, FishingForPlasmids outperformed the rest of the tools and only included chromosomal sequences in 7 (1.8%) bins. The rest of the tools included chromosomal sequences in a range from 25.25% to 51.73% of the bins. Surprisingly, MOB-suite included chromosomal sequences in 297 (40.2%) bins, including 65 chromosome-only predictions (chromosome contamination = 1).

A fraction of the plasmids ( $n = 28$ , 4.4%) were completely absent (recall = 0) from contig sequences and nodes in the assembly graph. Interestingly, 14 of these replicons were correctly reconstructed by plasmidSPAdes when using pair-end reads as input. This suggests that the quality of the assembly has impacted the ability of the tools to reconstruct certain plasmids. Consequently, it is possible that plasmid predictions for *E. coli* could be optimized by running SPAdes with different parameters, by performing assembly with different assemblers or through construction of Illumina libraries with a different read length.

The results from our study indicate that accurate reconstruction of *E. coli* plasmids from short-reads is still challenging using currently available bioinformatic methods. Long reads generated by Oxford Nanopore or PacBio technologies can span repeat elements in the bacterial genomes and are therefore use-

ful to obtain complete plasmid sequences. However, long-reads still exhibit a lower sequencing accuracy than Illumina reads [56], and small plasmids (size < 10 kb) are frequently underrepresented or absent in Nanopore libraries [57,58]. Consequently, combining long- and short-read sequences is currently the best option for correctly reconstructing *E. coli* plasmids. Nevertheless, the accuracy of long-reads has been increasing in recent years, mainly due to the release of improved hardware and also owing to the development of bioinformatic tools designed for read error correction [56]. It is possible that in the near future long-read only assemblies will provide the best alternative for obtaining complete bacterial genomes.

Nonetheless, in the absence of long-reads, bioinformatic tools can be applied to gain valuable insight on different aspects of the plasmidome of *E. coli*. MOB-suite presented the best overall performance of all tools, but predictions were frequently contaminated with chromosomal sequences. Consequently, using MOB-suite coupled to a binary classification tool could improve plasmid predictions in *E. coli*. Furthermore, these predictions could be used as an initial screening step for selecting interesting isolates for long-read sequencing.

## References

1. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012, 486, 207–214.
2. Ochoa, T.J.; Barletta, F.; Contreras, C.; Mercado, E. New insights into the epidemiology of enteropathogenic *Escherichia coli* infection. *Trans. R. Soc. Trop. Med. Hyg.* 2008, 102, 852–856.
3. Biran, D.; Ron, E.Z. Extraintestinal Pathogenic *Escherichia coli*. In *Current Topics in Microbiology and Immunology*; Springer: Cham, Switzerland, 2018; Volume 416, pp. 149–161.
4. European Centre for Disease Prevention and Control. Healthcare-associated infections acquired in intensive care units. In *ECDC Annual Epidemiological Report for 2017*; ECDC: Stockholm, Sweden, 2019.
5. Laupland, K.B.; Church, D.L. Population-Based Epidemiology and Microbiology of Community-Onset Bloodstream Infections. *Clin. Microbiol. Rev.* 2014, 27, 647–664.
6. Denamur, E.; Clermont, O.; Bonacorsi, S.; Gordon, D. The population genetics of pathogenic *Escherichia coli*. *Nat. Rev. Genet.* 2021, 19, 37–54.
7. Cassini, A.; Högberg, L.D.; Plachouras, D.; Quattrocchi, A.; Hoxha, A.; Simonsen, G.S.; Colomb-Cotinat, M.; Kretzschmar, M.E.; Devleeschauwer, B.; Cecchini, M.; et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: A population-level modelling analysis. *Lancet Infect. Dis.* 2019, 19, 56–66.
8. Stephens, C.; Arismendi, T.; Wright, M.; Hartman, A.; Gonzalez, A.; Gill, M.; Pandori, M.; Hess, D. F Plasmids Are the Major Carriers of Antibiotic Resistance Genes in Human-Associated Commensal *Escherichia coli*. *mSphere* 2020, 5.
9. Matamoros, S.; Van Hattem, J.M.; Arcilla, M.S.; Willemse, N.; Melles, D.C.; Penders, J.; Vinh, T.N.; Hoa, N.T.; Bootsma, M.C.J.; Van Genderen, P.J.; et al. Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the *mcr-1* gene indicates bacterial diversity but plasmid restriction. *Sci. Rep.* 2017, 7, 1–9.
10. Kaper, J.B. Pathogenic *Escherichia coli*. *Int. J. Med. Microbiol.* 2005, 295, 355–356.
11. Arredondo-Alonso, S.; Willems, R.; van Schaik, W.; Schürch, A.C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genom.* 2017, 3, e000128.
12. Wick, R.R.; Judd, L.; Gorrie, C.; Holt, K. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* 2017, 3, e000132.
13. Wick, R.R.; Judd, L.M.; Gorrie, C.L.; Holt, K.E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* 2017, 13, e1005595.
14. Carattoli, A.; Zankari, E.; García-Fernández, A.; Larsen, M.V.; Lund, O.; Villa, L.; Aarestrup, F.; Hasman, H. *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob. Agents Chemother.* 2014, 58, 3895–3903.
15. Ten Doesschate, T.; Abbott, I.J.; Willems, R.J.L.; Top, J.; Rogers, M.R.; Bonten, M.M.; Paganelli, F.L. In vivo acquisition of fosfomycin resistance in *Escherichia coli* by *fosA* transmission from commensal flora. *J. Antimicrob. Chemother.* 2019, 74, 3630–3632.
16. Gan, H.M.; Eng, W.W.H.; Dhanoa, A. First genomic insights into carbapenem-resistant *Klebsiella pneumoniae* from Malaysia. *J. Glob. Antimicrob. Resist.* 2020, 20, 153–159.
17. Van Driessche, L.; Vanneste, K.; Bogaerts, B.; De Keersmaecker, S.C.; Roosens, N.H.; Haesebrouck, F.; De Cremer, L.; Deprez, P.; Pardon, B.; Boyen, F. Isolation of Drug-Resistant *Gallibacterium anatis* from Calves with Unresponsive Bronchopneumonia, Belgium. *Emerg. Infect. Dis.* 2020, 26, 721–730.
18. Gupta, S.K.; Shin, H.; Han, D.; Hur, H.G.; Uno, T. Metagenomic analysis reveals the prevalence and persistence of antibiotic- and heavy metal-resistance genes in wastewater treatment plant. *J. Microbiol.* 2018, 56, 408–415.

19. Arredondo-Alonso, S.; Top, J.; McNally, A.; Puranen, S.; Pesonen, M.; Pensar, J.; Marttinen, P.; Braat, J.C.; Rogers, M.R.C.; van Schaik, W.; et al. Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*. *mBio* 2020, 11.
20. Arredondo-Alonso, S.; Top, J.; Corander, J.; Willems, R.J.; Schürch, A.C. Mode and dynamics of vanA-type vancomycin-resistance dissemination in Dutch hospitals. *Genome Med.* 2020, 13, 1–8.
21. Kans, J. Entrez Direct: E-utilities on the Unix Command Line. In *Entrez Programming Utilities Help* [Internet]; 2013 Apr 23 [Updated 2021 Jul 16]; National Center for Biotechnology Information: Bethesda, MD, USA, 2010.
22. Beghain, J.; Bridier-Nahmias, A.; Le Nagard, H.; Denamur, E.; Clermont, O. ClermonTyping: An easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb. Genom.* 2018, 4, e000192.
23. Lees, J.A.; Harris, S.R.; Tonkin-Hill, G.; Gladstone, R.A.; Lo, S.W.; Weiser, J.N.; Corander, J.; Bentley, S.D.; Croucher, N.J. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* 2019, 29, 304–316.
24. Argimón, S.; Abudahab, K.; Goater, R.J.E.; Fedosejev, A.; Bhai, J.; Glasner, C.; Feil, E.J.; Holden, M.T.G.; Yeats, C.A.; Grundmann, H.; et al. Microreact: Visualizing and sharing data for genomic epidemiology and phylo-geography. *Microb. Genom.* 2016, 2, e000093.
25. Walk, S.T.; Alm, E.W.; Gordon, D.M.; Ram, J.L.; Toranzos, G.A.; Tiedje, J.M.; Whittam, T.S. Cryptic Lineages of the Genus *Escherichia*. *Appl. Environ. Microbiol.* 2009, 75, 6534–6544.
26. Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinform.* 2020, 70, e102.
27. Royer, G.; Decousser, J.W.; Branger, C.; Dubois, M.; Médigue, C.; Denamur, E.; Vallenet, D. PlaScope: A targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb. Genom.* 2018, 4.
28. Schwengers, O.; Barth, P.; Falgenhauer, L.; Hain, T.; Chakraborty, T.; Goesmann, A. Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb. Genom.* 2020, 6, e000398.
29. Arredondo-Alonso, S.; Rogers, M.R.C.; Braat, J.C.; Verschuuren, T.D.; Top, J.; Corander, J.; Willems, R.J.L.; Schürch, A.C. mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genom.* 2018, 4, e000224.
30. Krawczyk, P.S.; Lipinski, L.; Dziembowski, A. PlasFlow: Predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* 2018, 46, e35.
31. Pellow, D.; Mizrahi, I.; Shamir, R. PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.* 2020, 16, e1007781.
32. van Bloois, L.V.D.G.; Wagenaar, J.A.; Zomer, A.L. RFPlasmid: Predicting plasmid sequences from short read assembly data using machine learning. *bioRxiv* 2020.
33. Fang, Z.; Tan, J.; Wu, S.; Li, M.; Xu, C.; Xie, Z.; Zhu, H. PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* 2019, 8.
34. Antipov, D.; Hartwick, N.; Shen, M.; Rayko, M.; Lapidus, A.; Pevzner, P.A. plasmidSPAdes: Assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016, 32, 3380–3387.
35. Rozov, R.; Brown Kav, A.; Bogumil, D.; Shterzer, N.; Halperin, E.; Mizrahi, I.; Shamir, R. Recycler: An algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 2017, 33, 475–482.
36. Arredondo-Alonso, S.; Bootsma, M.; Hein, Y.; Rogers, M.R.C.; Corander, J.; Willems, R.J.L.; Schürch, A.C. gplas: A comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics* 2020, 36, 3874–3876.
37. Robertson, J.; Nash, J.H.E. MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* 2018, 4, e000206.



38. Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016, 17, 132.
39. Carattoli, A.; Hasman, H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *Methods Mol. Biol.* 2020, 2075, 285–294.
40. Pellow, D.; Zorea, A.; Probst, M.; Furman, O.; Segal, A.; Mizrahi, I.; Shamir, R. SCAPP: An algorithm for improved plasmid assembly in metagenomes. *Microbiome* 2021, 9, 1–12.
41. Goswami, C.; Fox, S.; Holden, M.T.; Connor, M.; Leanord, A.; Evans, T.J. Origin, maintenance and spread of antibiotic resistance genes within plasmids and chromosomes of bloodstream isolates of *Escherichia coli*. *Microb. Genom.* 2020, 6, mgen000353.
42. Peter, S.; Bosio, M.; Gross, C.; Bezdán, D.; Gutierrez, J.; Oberhettinger, P.; Liese, J.; Vogel, W.; Dörfel, D.; Berger, L.; et al. Tracking of Antibiotic Resistance Transfer and Rapid Plasmid Evolution in a Hospital Setting by Nanopore Sequencing. *mSphere* 2020, 5.
43. Gong, L.; Tang, N.; Chen, D.; Sun, K.; Lan, R.; Zhang, W.; Zhou, H.; Yuan, M.; Chen, X.; Zhao, X.; et al. A Nosocomial Respiratory Infection Outbreak of Carbapenem-Resistant ST131 With Multiple Transmissible Carrying Plasmids. *Front. Microbiol.* 2020, 11, 2068.
44. Paramita, R.I.; Nelwan, E.J.; Fadilah, F.; Renesteen, E.; Puspadari, N.; Erlina, L. Genome-based characterization of *Escherichia coli* causing bloodstream infection through next-generation sequencing. *PLoS ONE* 2020, 15, e0244358.
45. Mughini-Gras, L.; Dorado-García, A.; van Duijkeren, E.; van den Bunt, G.; Dierikx, C.M.; Bonten, M.J.M.; Bootsma, M.C.J.; Schmitt, H.; Hald, T.; Evers, E.G.; et al. Attributable sources of community-acquired carriage of *Escherichia coli* containing  $\beta$ -lactam antibiotic resistance genes: A population-based modelling study. *Lancet Planet Health* 2019, 3, e357–e369.
46. Ludden, C.; Raven, K.E.; Jamrozny, D.; Gouliouris, T.; Blane, B.; Coll, F.; de Goffau, M.; Naydenova, P.; Horner, C.; HernandezGarcia, J.; et al. One Health Genomic Surveillance of *Escherichia coli* Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *mBio* 2019, 10, e02693-18.
47. Hendriksen, R.S.; Bortolaia, V.; Tate, H.; Tyson, G.H.; Aarestrup, F.; McDermott, P.F. Using Genomics to Track Global Antimicrobial Resistance. *Front. Public Health* 2019, 7, 242.
48. Rodríguez-Beltrán, J.; DelaFuente, J.; León-Sampedro, R.; MacLean, R.C.; Millán, Á.S. Beyond horizontal gene transfer: The role of plasmids in bacterial evolution. *Nat. Rev. Genet.* 2021, 19, 347–359.
49. Che, Y.; Yang, Y.; Xu, X.; Břinda, K.; Polz, M.F.; Hanage, W.P.; Zhang, T. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc. Natl. Acad. Sci. USA* 2021, 118.
50. He, S.; Hickman, A.B.; Varani, A.M.; Siguier, P.; Chandler, M.; Dekker, J.P.; Dyda, F. Insertion Sequence IS 26 Reorganizes Plasmids in Clinically Isolated Multidrug-Resistant Bacteria by Replicative Transposition. *mBio* 2015, 6, e00762-15.
51. 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.
52. Shaw, L.P.; Chau, K.K.; Kavanagh, J.; AbuOun, M.; Stubberfield, E.; Gweon, H.S.; Barker, L.; Rodger, G.; Bowes, M.J.; Hubbard, A.; et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci. Adv.* 2021, 7, eabe3868.
53. Pesesky, M.W.; Tilley, R.; Beck, D. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid* 2019, 102, 10–18.
54. Jesus, T.F.; Ribeiro-Gonçalves, B.; Silva, D.N.; Bortolaia, V.; Ramirez, M.; Carriço, J.A. Plasmid ATLAS: Plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.* 2019, 47, D188–D194.

55. Bosi, E.; Fani, R.; Fondi, M. The mosaicism of plasmids revealed by atypical genes detection and analysis. *BMC Genom.* 2011, 12, 403.
56. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020, 21, 1–16.
57. Arredondo-Alonso, S.; Pöntinen, A.K.; Cléon, F.; Gladstone, R.A.; Schurch, A.C.; Johnsen, P.J.; Samuelson, O.; Corander, J. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *BioRxiv* 2021, 6, 448320.
58. Wick, R.R.; Judd, L.M.; Wyres, K.L.; Holt, K.E. Recovery of small plasmid sequences via Oxford Nanopore sequencing. *BioRxiv* 2021.

## Supplementary Materials

### Contig Alignments

For determining the composition of the bins, we used QUAST (v.5.0.2) to map the contigs of each bin to the corresponding closed reference genome. Quast uses nucmer for contig alignment. We only took contigs with a length larger than 1kb into account. Alignment of a contig was considered as correct when the sequence identity was greater than 95% and the query coverage was more than 90%, even if it was classified as “relocation” or “inversion” by QUAST (Table S4). Contigs that ambiguously aligned to several replicons were considered as correct, given that they met the previous conditions.

The alignment length of cases that were classified as ‘translocation’ by QUAST, specifying a sequence where the left and right flanking regions map to different replicons, were computed separately for all replicons.

### Maximum theoretical recall for plasmid reconstruction

To determine the maximum recall for plasmid reconstruction from the input data (assembled short read contigs or assembly graph), we aligned either contigs or nodes extracted from the assembly graph to their respective closed reference genomes. This was important for identifying plasmid fragments that could have been missed when sequencing and which could consequently be absent in the final assembly (dead-end in the assembly graph). Therefore, these fragments could never have been reconstructed by the tools that use assembled genomes as input. It also revealed potential mismatches when mapping to the reference sequence.

Plasmids sequences were recovered with a median recall of 0.97 (IQR=0.08) when using contigs as input, and a median recall of 0.95 (IQR=0.13) was obtained when using nodes in the assembly graphs for the alignment (Figure S5 A). Notably, a total of 185 plasmids (29,3%) were perfectly recovered from contigs (recall=1), the majority of which (n=143, 77%) were small sized plasmids presenting lengths below 18 kbp. Similarly, 139 (22%) plasmids were fully recovered from assembly graphs, these were mostly small plasmids (n=122, 88%). Furthermore, 53 plasmids were fully recovered from contigs sequences only, while 7 were solely extracted from nodes in the assembly graph (Figure S5 B and C, Table S5).

Interestingly, we found that 31 (4.9%) plasmid sequences were completely missing from contigs sequences (recall=0), while 32 (5%) were missing from the assembly graph. A total of 28 plasmids (4.4%) were absent from both types of input (Figure S5 D). The sizes of these missing plasmids ranged from 763 to 4087 bp and did not contain any antibiotic resistant determinants. Interestingly, two small plasmids (Accessions: CP049974.1 and CP057228.1) were completely assembled from contigs (recall=1) but missing from the assembly graph produced by SPAdes (Table S5).

Finally, we discovered that the majority of small plasmids were contained in single nodes in the assembly graph (n=215, 79.05%) or single contigs (n=220,80.88%). Additionally, we found that 14 (5.15%) small plasmids were contained in hybrid contigs, formed by sequences derived from more than one replicon (Table S5).

## Supplementary Data

Supplementary data can be downloaded from:

<https://www.mdpi.com/article/10.3390/microorganisms9081613/s1>.

## Supplementary Figures

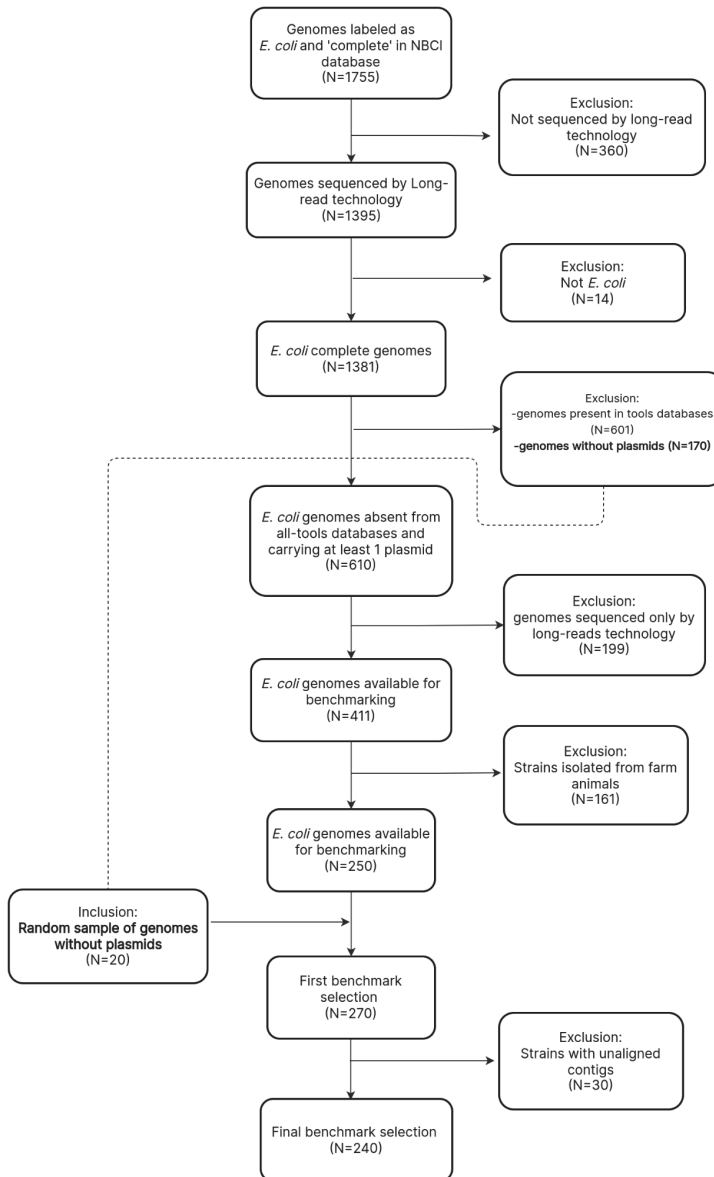
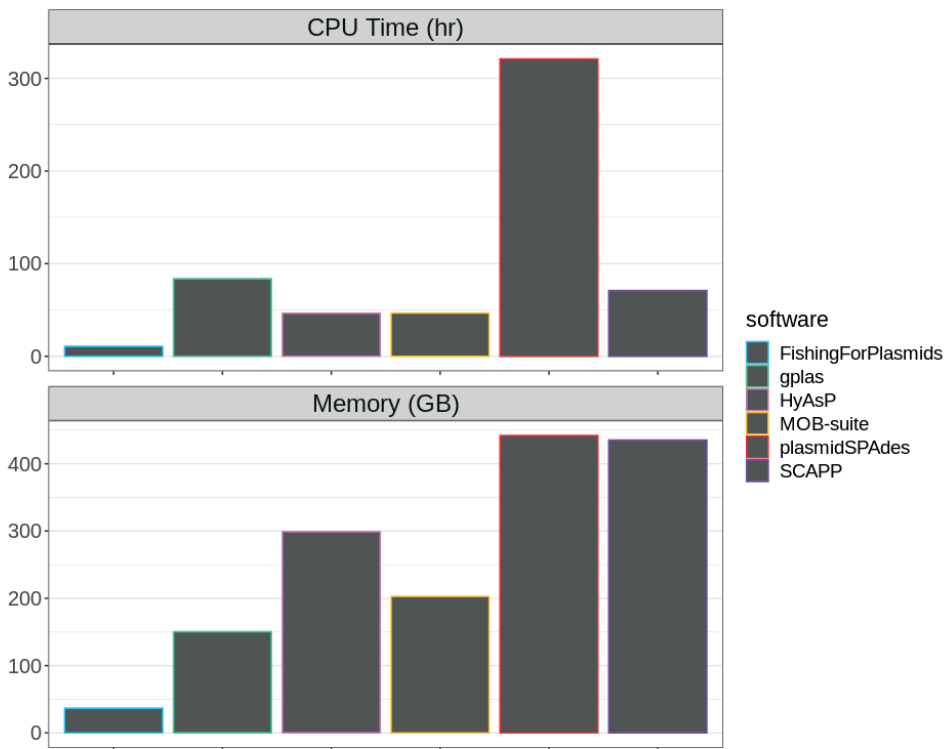
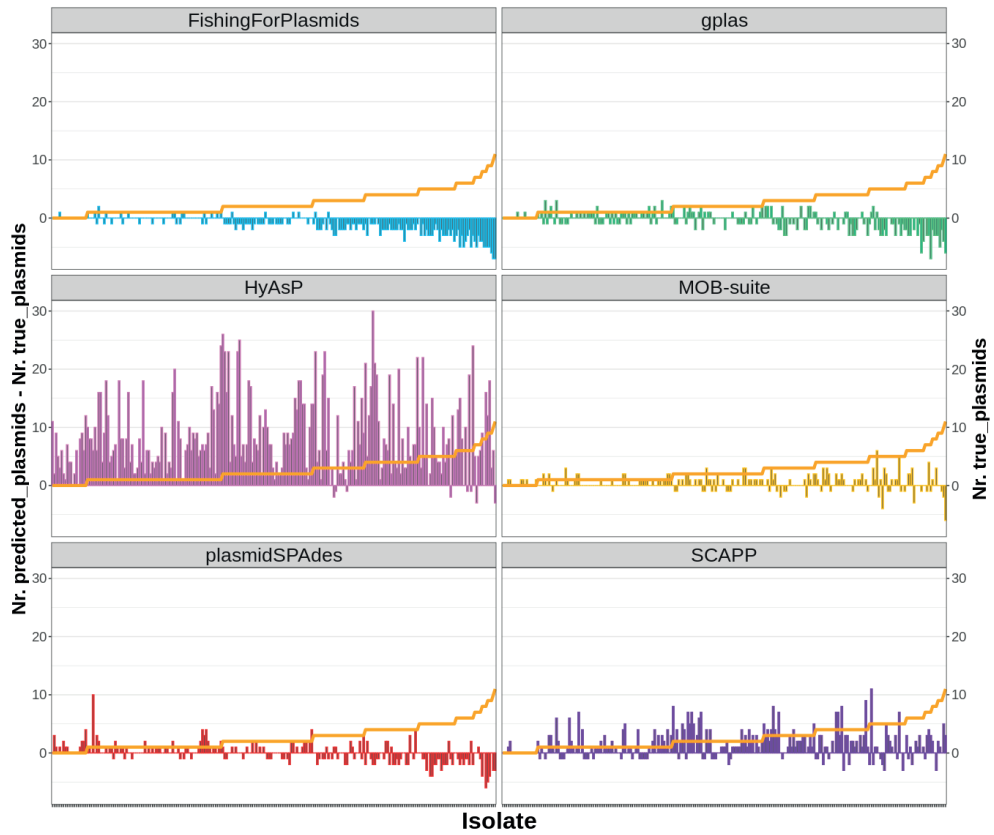


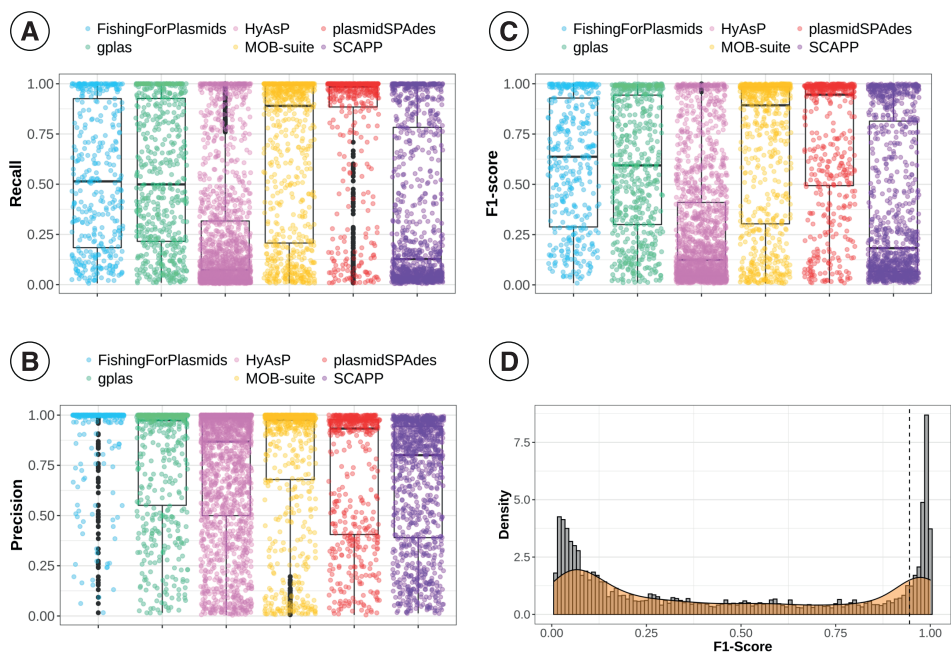
Figure S1. This flowchart summarizes the steps applied for selecting the 240 *E. coli* sequences included in this benchmark.



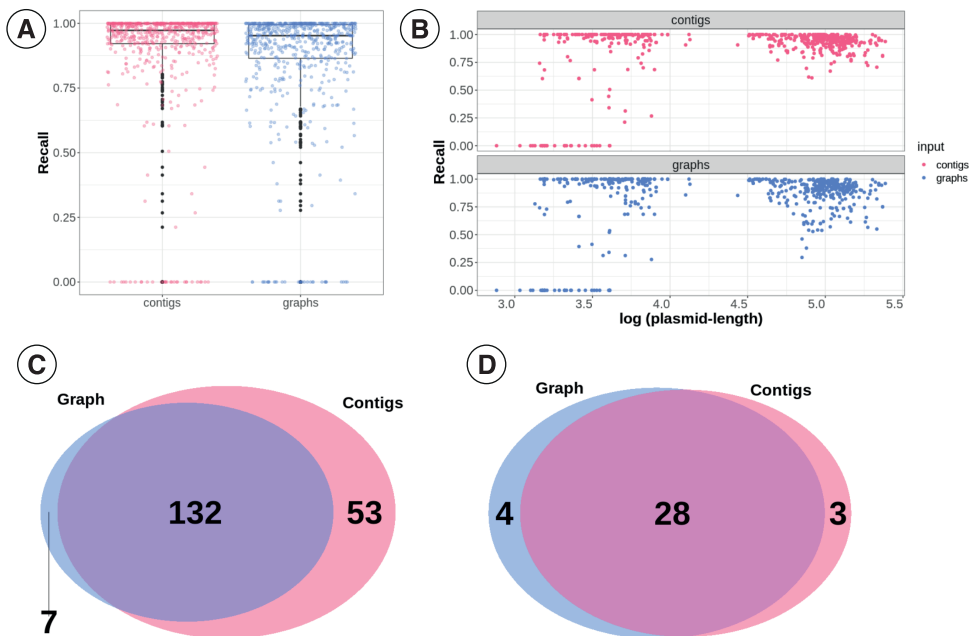
**Figure S2.** Total CPU-Time (top) and Memory (bottom) required by each tool to generate plasmid predictions in 270 *E. coli* genomes. Only 240 of these genomes were included in the final benchmark (see Methods).



**Figure S3.** Difference between predicted and observed plasmid content per tool. Strains were ordered by increasing amount of plasmids (orange line, right y-axis). A negative value on the left y-axis indicates an underestimation of the amount of plasmids predicted for that particular strain whereas a positive value indicates an overestimation.

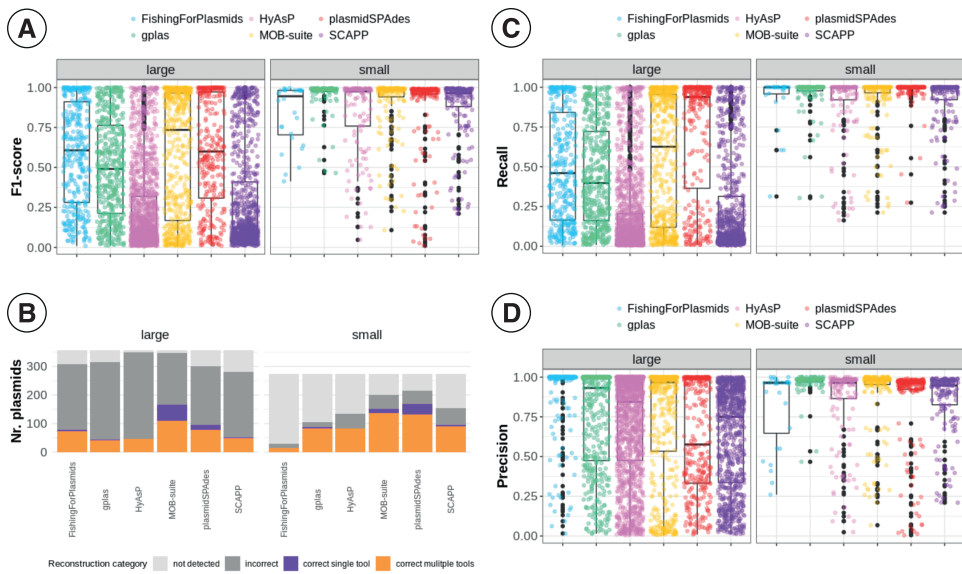


**Figure S4.** Recall (A), Precision (B) and F1-score (C) values distribution for all plasmid predictions made by each tool. (D) F1-score distribution of all plasmid predictions made by all tools combined. We established an F1-score cut-off of 0.95 (dashed line) to define a plasmid prediction as correct.

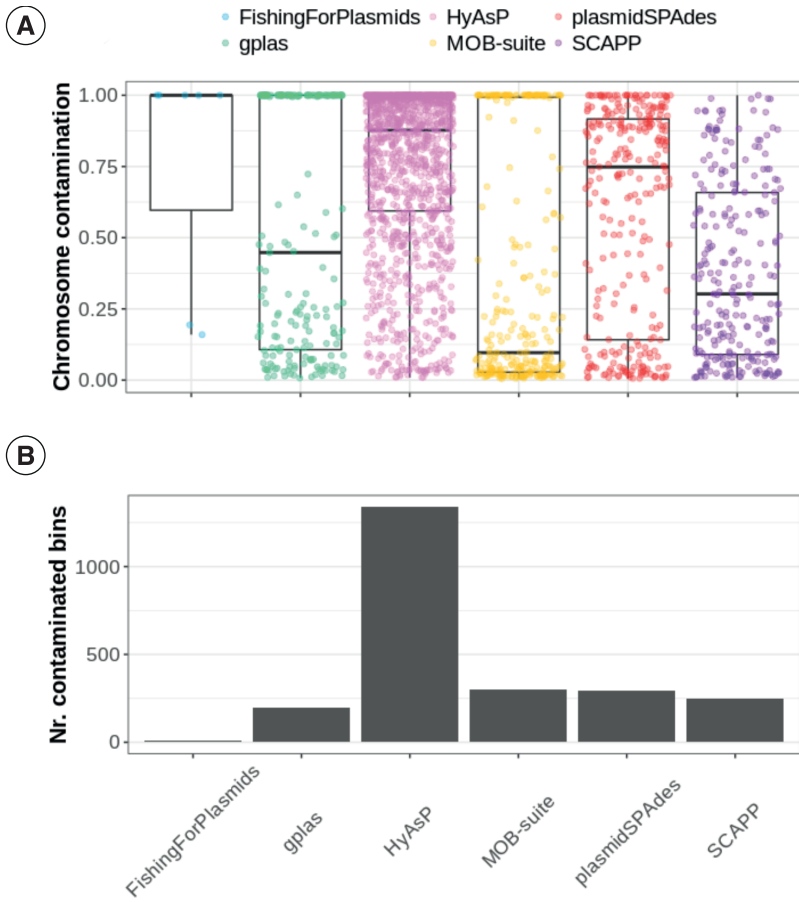


**Figure S5.** (A) Recall values obtained when aligning all assembled contigs or nodes in the assembly graph to reference plasmids. (B) Same recall values as A, as a function of plasmid size. (C) Venn diagram that shows the number of fully recovered plasmids (recall=1). (D) Venn diagram that shows completely missed plasmids (recall=0).

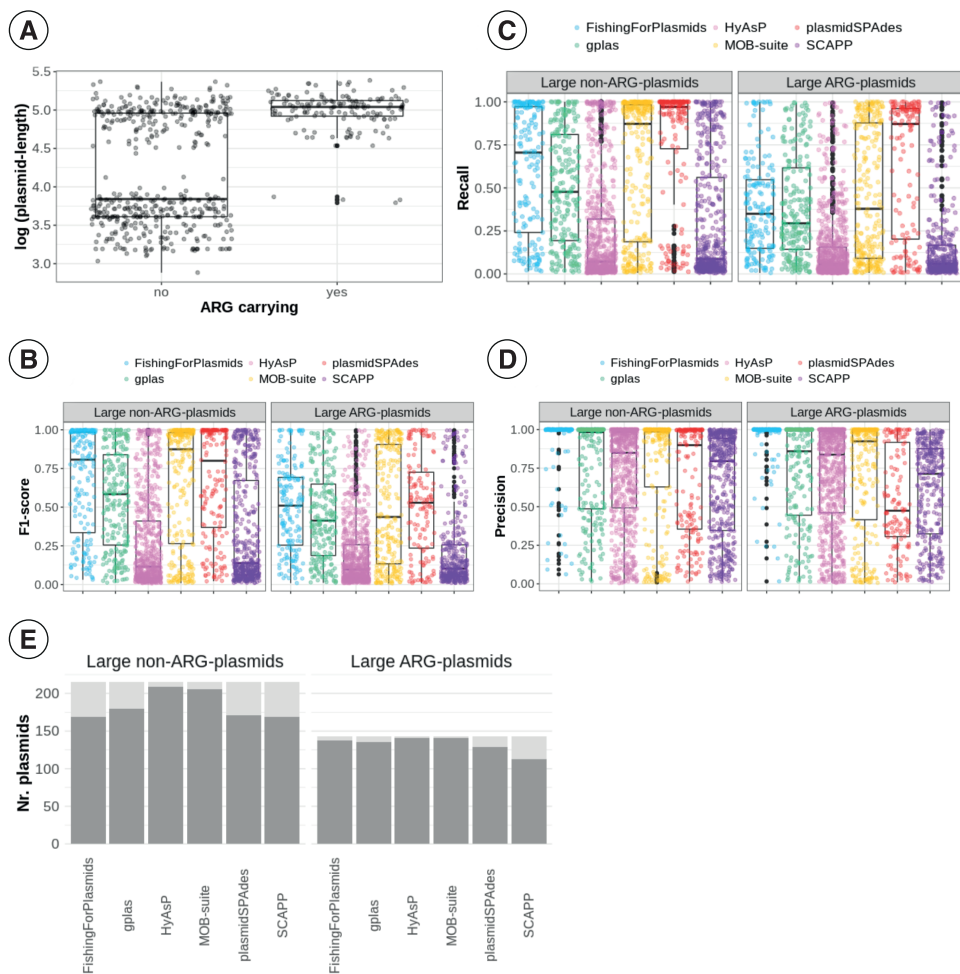




**Figure S6.** (A) F1-score value distribution for plasmid predictions according to plasmid sizes. (B) Absolute count of small and large reference plasmids that were correctly reconstructed (F1-score  $\geq 0.95$ ) by a single or multiple tools, incorrectly reconstructed (F1-score  $< 0.95$ ) and not detected. (C) Recall value distribution for plasmid predictions according to plasmid sizes. (D) Precision value distribution for plasmid predictions according to plasmid sizes.



**Figure S7.** (A) Distribution of chromosome contamination values per tool. Each dot corresponds to an individual prediction that presented chromosomal sequences. (B) Count of predictions that contained chromosomal contamination.



**Figure S8.** (A) Distribution of lengths for ARG and non-ARG plasmids. (B) F1-score values distribution per tool for predictions of large ARG plasmids vs. large non-ARG plasmids. (C) Recall values distribution per tool for predictions of large ARG plasmids vs. large non-ARG plasmids. (D) Precision values distribution per tool for predictions of large ARG plasmids vs. large non-ARG plasmids. (E) Bar plots showing absolute counts of detected and not detected reference plasmids.

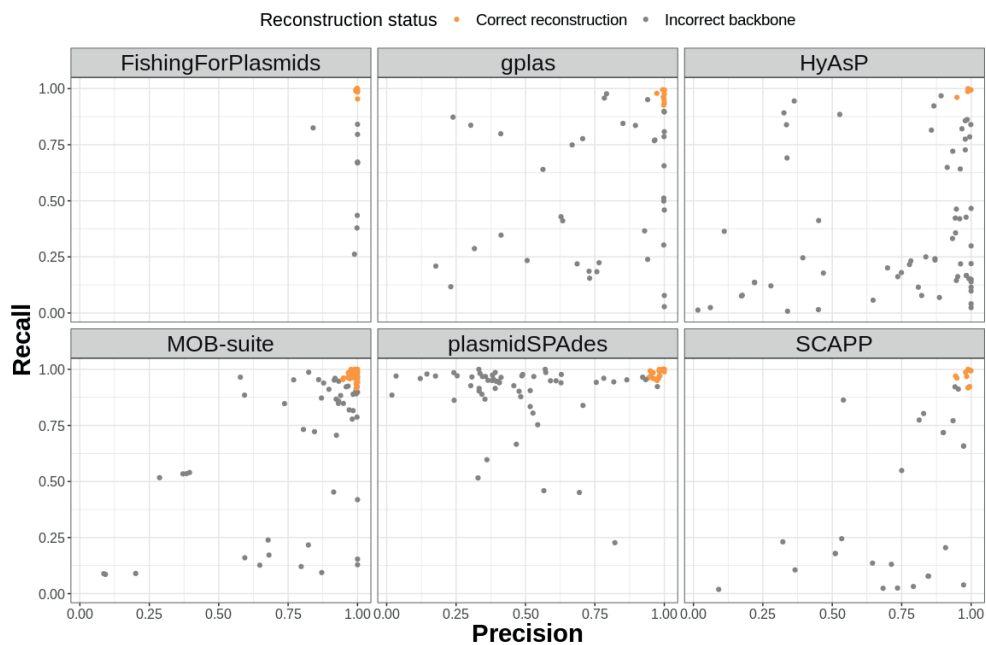
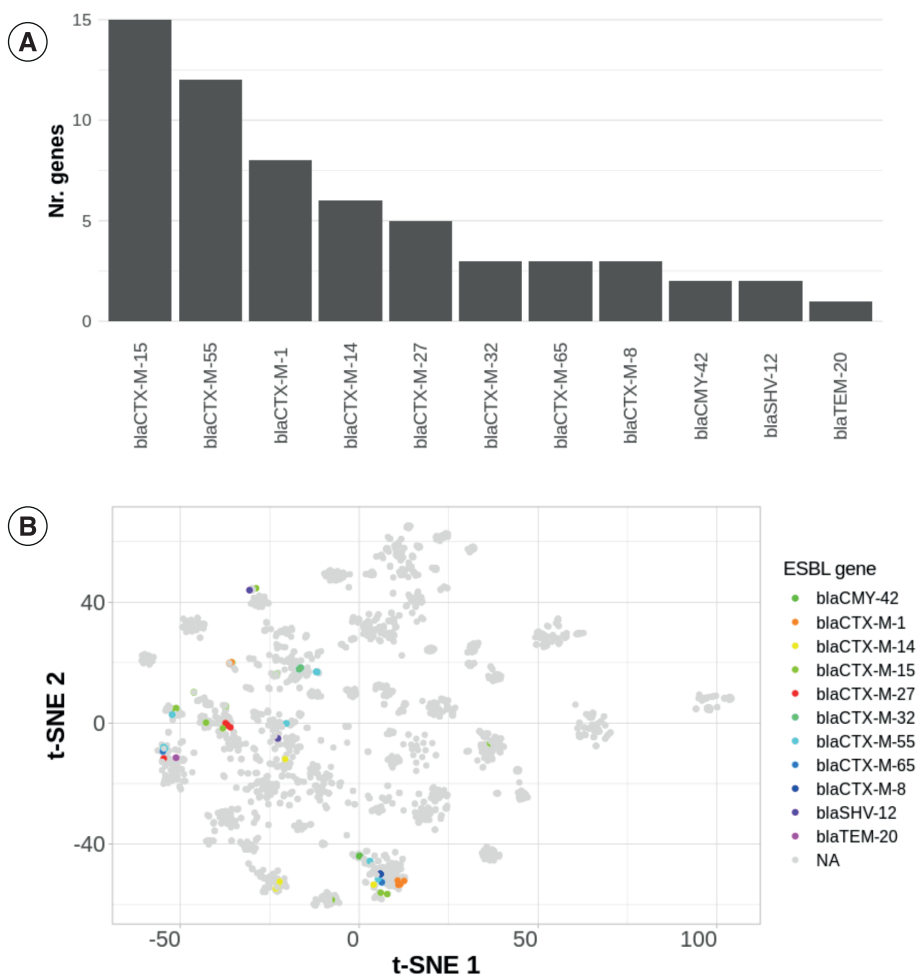
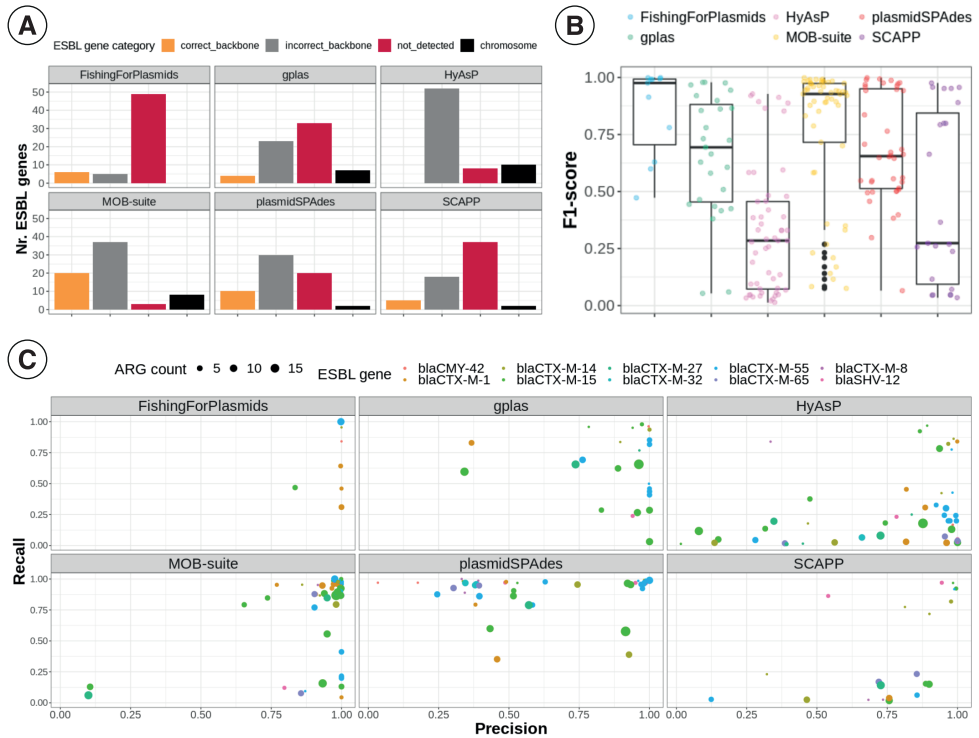


Figure S9. Scatter plot that shows precision (bp) and recall (bp) values for predictions that presented a Recall(ARG) equal to 1.

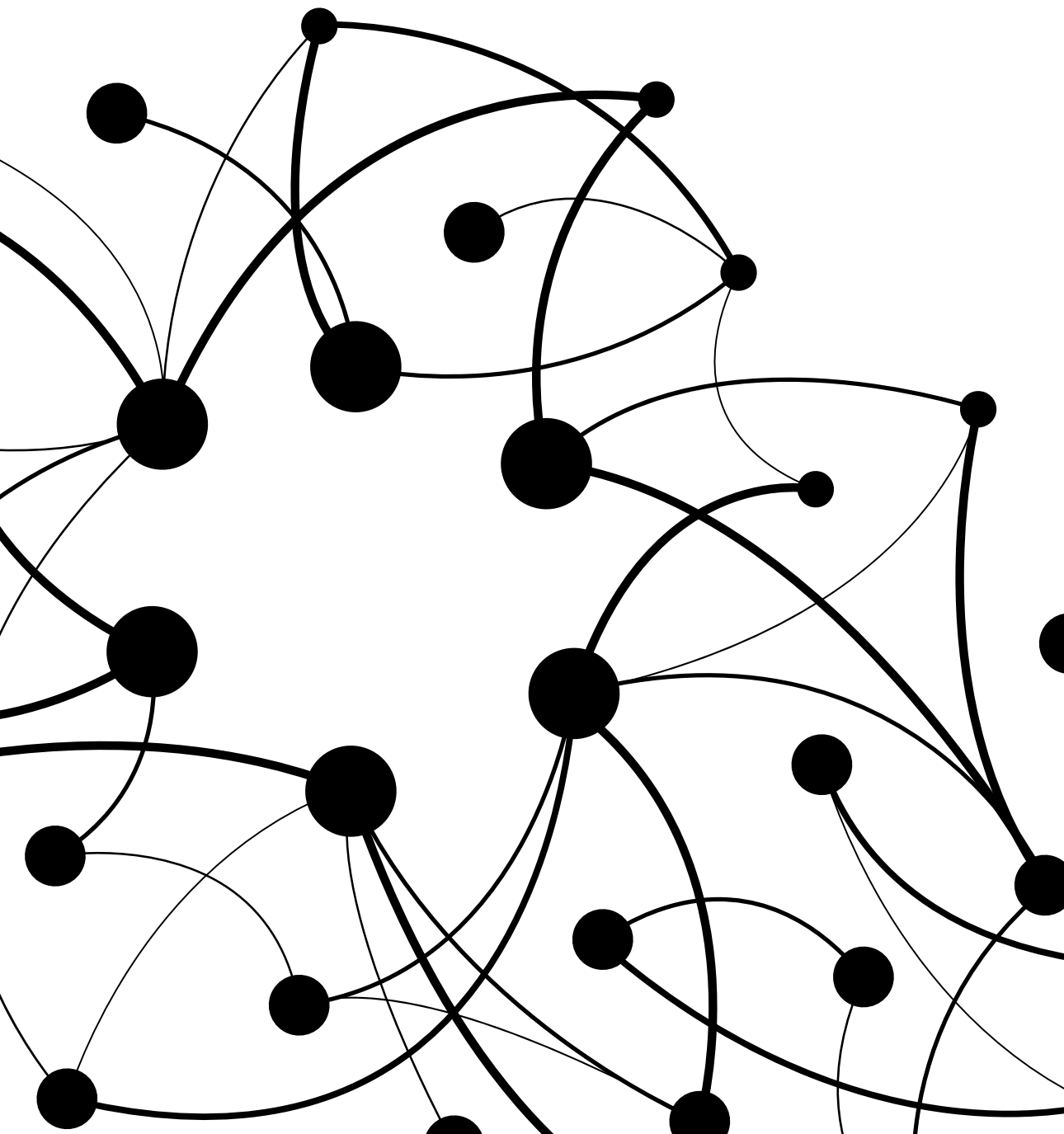


**Figure S10.** (A) Absolute count of ESBL variants in the benchmark data set. (B) tSNE created based on plasmids k-mer distances obtained with Mash ( $k=21$ ,  $s1000$ ). ESBL-plasmids included in the benchmark are colored according to distinct ESBL genes.



**Figure S11.** (A) Absolute count of ESBL genes according to prediction status. The different prediction status were determined according to the following criteria. Correct backbone: ESBL gene was included in a bin that presented an F1-Score  $\geq 0.95$ , incorrect backbone: ESBL gene was included in a bin that presented an F1-Score  $< 0.95$ , not detected: ESBL gene was not included in the bins produced by the tool, chromosome: chromosome-derived ESBL gene was included in the bins generated by the tool. (B) F1-score value distribution for all bins containing plasmid-derived ESBL genes. (C) Precision vs Recall plot for all bins containing a plasmid-derived ESBL gene.







# 03

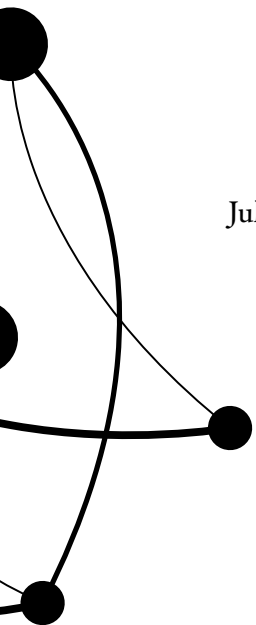
---

## **An optimised short-read approach to predict and reconstruct antibiotic resistance plasmids in *Escherichia coli***

Julian A. Paganini, Lisa Vader, Jesse J. Kerkvliet, Nienke L. Plantinga,  
Rodrigo Meneses, Jukka Corander, Rob J.L. Willems,  
Sergio Arredondo-Alonso\* and Anita C. Schürch\*

\*Authors contributed equally

Manuscript in preparation



### Abstract

*Escherichia coli* has become the most prevalent resistant pathogen worldwide, being responsible for more than 250,000 deaths each year. Antibiotic resistance genes (ARGs) in *E. coli* are frequently encoded by plasmids, mobile genetic elements that play a pivotal role in the spread of resistance. Accurate reconstruction of *E. coli* 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 an ensemble classifier, named plasmidEC, that identifies plasmid-derived contigs by combining the output of three binary classification tools. We demonstrated that plasmidEC is especially suited to classify contigs derived from ARG plasmids (recall of 0.941). Subsequently, we optimised gplas, a graph-based tool that bins plasmid-predicted contigs into distinct plasmid predictions. The new version of gplas is more effective at recovering plasmids with large sequencing coverage variations and can be combined with the output of any binary classification tool. The combination of plasmidEC with gplas showed a high completeness (median=0.818) and F1-score (median=0.812) when reconstructing ARG plasmids, which exceeded the binning capacity of the reference-based method MOB-suite. In the absence of long-read data, our method offers the best alternative to reconstruct ARG plasmids in *E. coli*.

### Data Summary

No new sequencing data has been generated in this study. All genomes used in this research are publicly available at the GenBank and Sequence Read Archive of the National Center for Biotechnology Information. Accession numbers are specified in Supplementary Table S1.

Scripts to reproduce the results reported in this manuscript 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>, and gplas can be found at <https://gitlab.com/mmb-umcu/gplas>.

## Introduction

*Escherichia coli* is a commensal gram-negative bacterium inhabiting the gastro-intestinal 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 (MDR) *E. coli* lineages has narrowed down the treatment options for infections with these bacteria [3,4]. Moreover, a recent assessment of the global burden of antimicrobial resistance (AMR) estimated that AMR *E. coli* infections accounted for more than 250,000 deaths in 2019, placing *E. coli* amongst the most prevalent AMR pathogens worldwide [5].

Horizontal gene transfer (HGT) is one of the main drivers behind the rapid spread of AMR [6–8]. AMR genes are commonly associated with mobile genetic elements (MGEs), such as transposons, integrative and conjugative elements, and plasmids, which facilitate their mobility across bacteria [9,10]. Of these MGEs, plasmids play a pivotal role by disseminating AMR genes 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 by 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 tangled with other genomic elements (chromosomal sequences). This makes it hard to reconstruct complete plasmids from short-read sequencing data [18].

Several bioinformatic tools are currently available to predict plasmids from short-read sequencing data. They can be broadly categorised into two groups: (i) tools that produce a binary classification of contigs as either plasmid- or chromosome-derived, generating an output that predicts 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, provide a more suitable output for plasmid epidemiology studies.

We recently evaluated the performance of several plasmid reconstruction tools for use with *E. coli* [19]. We found that the best performing tool, MOB-suite [20], only achieved the correct reconstruction of 50.2% of the plasmids. Moreover, all tools performed sub-optimally when attempting to reconstruct plasmids containing antibiotic resistance genes (ARG-plasmids), ranging from 3.4% to 27.9% correct ARG-plasmid reconstructions. These results emphasised 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. Firstly, we optimised gplas, a plasmid binning tool, to recover walks in the assembly graph corresponding to plasmids with a pronounced coverage variation. Secondly, we developed an ensemble classifier, plasmidEC, combining multiple existing binary classification tools (Plascope [21], RFPplasmid [22], Platon [23] and mlplasmids [24]) to predict plasmid-derived contigs. Thirdly,

we coupled plasmidEC with gplas to accurately bin plasmid-derived contigs into separate components corresponding to individual plasmid sequences. Comparisons with existing methods revealed that our method outperformed all currently available plasmid reconstruction tools for *E. coli*, 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 *E. coli* complete genomes from 8 different phylogroups and 117 sequence types (STs), encoding 631 plasmids, was selected as previously described in Paganini et al. [19]. The *E. coli* genomes originated 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/kbclin/ncbi-genome-download>), respectively. Genomes present in the training datasets or reference databases of existing plasmid classification tools (mlplasmids, PlaScope, Platon and/or RFPlasmid) were removed (n=26). The remaining 214 complete genomes, encoding 542 plasmids, were used to benchmark the binary classifiers (Supplementary Data 1). From these, 15 genomes (Supplementary Data 2) were randomly selected to optimise 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 (v3.11.1) for genome assembly [25]. 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)[26]. Only contigs larger than 1,000 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 if they were exclusively aligned to either the chromosome or to plasmids (n=1,236). The same criterion was used for inclusion of misassembled contigs (n=1,862). In total, the benchmark dataset included 14,746 contigs (Supplementary Figure S1).

#### Assessment of the individual binary classifiers

Contigs were classified by mlplasmids (v2.1.0), PlaScope (v.1.3.1), Platon (v.1.6) and RFPlasmid (v0.0.17). 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 actual class of the contig, as described in the section above. For PlaScope, an ‘unclassified’ prediction was handled as a negative prediction. Predictions were categorised into: True Positives (TP, prediction = plasmid, class = plasmid), True Negatives (TN, prediction = chromosome, class = chromosome), False Positives (FP,

prediction = plasmid, class = chromosome) and False Negatives (FN, prediction = chromosome, class = plasmid). Global performance of the tools was evaluated with the following metrics:

$$\text{Recall (contig)} = \frac{TP}{TP + FN}$$

$$\text{Precision (contig)} = \frac{TP}{TP + FP}$$

$$F1 - \text{Score (contig)} = 2 \cdot \frac{\text{Recall (contig)} \times \text{Precision (contig)}}{\text{Recall (contig)} + \text{Precision (contig)}}$$

### 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/PlaScope/Platon, mlplasmids/PlaScope/RFPlasmid, mlplasmids/Platon/RFPlasmid and PlaScope/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

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

## **Benchmarking of 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 adapter contamination and bases with a phred quality score below 20. Unicycler (v0.4.8) [27] was then applied to perform *de novo* assembly with default parameters. Contigs larger than 1,000 bp were used as input for MOB-suite (v3.0.0) [20], while assembly graphs in GFA format served as input for gplas (v1.0.0). To run gplas, nodes from the graph were first classified as plasmid- or chromosome-derived using either plasmidEC or PlaScope; only nodes larger than 1,000 bp were classified. Output from the tools were modified to assign probabilities for the classification of each node, which is required by the gplas algorithm. For PlaScope, discrete probabilities were assigned based on the node classification status, if a node was classified as plasmid, a probability of 1 was assigned, while a 0 was appointed for chromosome-predicted nodes. In the case of unclassified nodes, a probability of 0.5 was designated. For plasmidEC, probabilities were based on the fraction of tools that agreed on the classification. For example, if 2 out of 3 tools agreed in classifying a node as plasmid, a probability of 0.66 was assigned.

### Analysis of the plasmid bin composition

To evaluate the bins (i.e. predicted individual plasmids) created by MOB-suite and gplas, we used QUAST (v5.0.2) [26] 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.

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

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

$$\text{F1 - Score (bp)} = \frac{2 \times \text{Accuracy (bp)} \times \text{Completeness (bp)}}{\text{Accuracy (bp)} + \text{Completeness (bp)}}$$

If a bin was composed of contigs derived from different plasmids, then accuracy, completeness and F1-score were reported for each plasmid-bin combination.

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

To determine combined recall for each reference plasmid, all bins generated in an isolate were combined as followed:

$$\text{Combined completeness (bp)} = \sum_1^n \text{Completeness (bp)}$$

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

### Antibiotic Resistance Gene (ARG) Prediction

Resistance genes were predicted by running Abricate (v1.0.1) against the Resfinder [28] database (database indexed on 19 April 2020) with reference plasmids as query, using 80% as identity and coverage cut-off. The same software and parameters were used to predict the presence of ARGs in the plasmid bins generated by each of the plasmid reconstruction tools.

### Evaluation of ARGs binning

For bins that carried ARGs, we calculated Recall(ARG) and Precision(ARG) as indicated below.

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

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

### 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 unbinned unitigs (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 recall that can be obtained to reconstruct every reference plasmid using short-read sequencing data. All nodes from the assembly graph, before applying any classification tool, were converted to FASTA format using the 'extract' option of gplas. Nodes smaller than 1,000 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 closed reference genomes using QUAST to obtain the recall values. This maximum recall value was called the recovered fraction (See supplementary results).

### Read coverage of missing reference plasmids

A small number of plasmids were either completely missed or recovered with low recall after short-read assembly. In order to determine if these sequences were also missing from short-reads, trimmed Illumina reads were aligned to reference genomes using BWA MEM (v.0.7.17) [29] with default parameters. Resulting SAM files were converted to BAM and sorted using SAMtools (v1.9) [30]. Read coverages per base were determined using BEDTOOLS (v2.30.0) [31] (See supplementary results).

## Results

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

The original gplas algorithm performs *de novo* reconstruction of plasmids through multiple steps (Figure 1 - Steps 1 to 3) [32]. 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 for this classification. 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 co-existence in the computed plasmid walks. A detailed description of the algorithm can be found in the original publication [32]. Given that gplas performed sub-optimally when reconstructing *E. coli* plasmids in our previous study [19], we improved the algorithm by introducing two major modifications:

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

Coupling gplas with an accurate binary classifier improves the reconstruction of plasmids, as we previously demonstrated for *Enterococcus faecalis* and *Klebsiella pneumoniae* [32,33]. Consequently, the gplas algorithm was adapted to accept predictions from any binary classifier.

### **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 (Figure 1 - Step 2). Consequently, plasmid-predicted unitigs that can't be connected to other unitigs through these paths are classified as unbinned, and are not included in the plasmid predictions (Figure 1- Step 3). Unbinned unitigs seem to originate from reference plasmids that were sequenced with a pronounced coverage variation (Supplementary Figure 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 (Figure 1 - Steps 4 & 5). Whenever unbinned unitigs are produced by the original gplas algorithm, gplas v1.0 will produce a second round of binning in bold mode by running two additional steps:

#### *1. 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 (Figure 3 - Step 4).

#### *2. Plasmidome network reconstruction and repartitioning.*

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

We optimised the predictions obtained with gplas v1.0 using a subset of 15 *E. coli* genomes that contained unbinned unitigs and that were excluded from subsequent benchmarking efforts (Supplementary Table 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 gplas v1.0 exhibited consistently higher completeness<sub>(bp)</sub> values when compared to the original predictions (Supplementary Figure S3 A). 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, 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. Nevertheless, this value can be adjusted by the user. As a single example, we display the plasmid predictions obtained with and without the incorporation of the bold mode for genome GCA\_013823335.1\_ASM1382333v1 (Supplementary Figure S3 B and S3 C). In this case, the bold walks allowed to recover 7 additional contigs belonging to plasmids CP057179.1 and CP057180.1.

Gplas v1.0, including the features discussed above and a detailed user guide, can be found at <https://gitlab.com/mmb-umcu/gplas>.



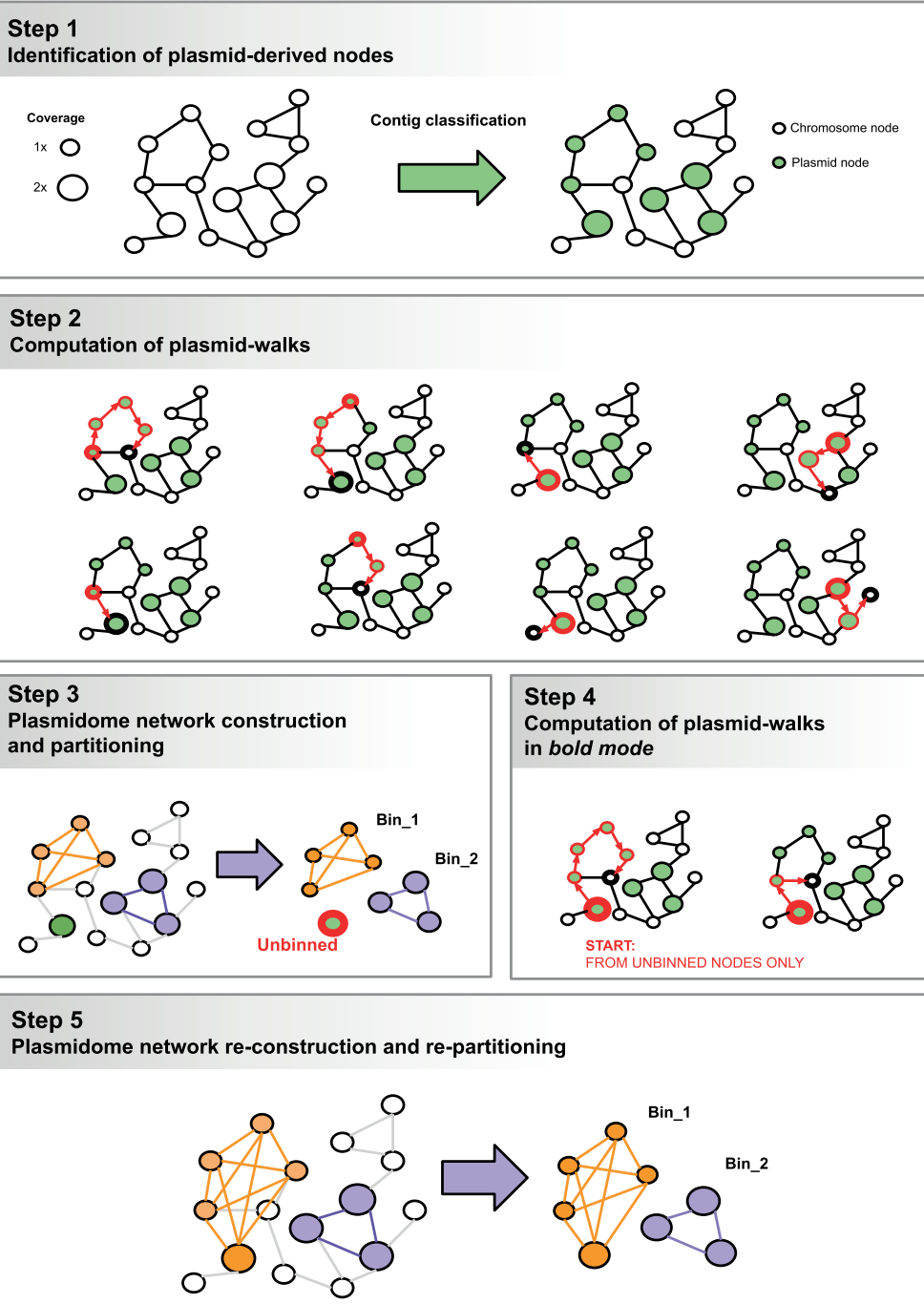


Figure 1. Schematics on gplas algorithm. The steps 4 and 5 were added to recover unbinned units.

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

In order to combine *gplas* with the best available binary classifier for *E. coli*, we compared the performance of four different tools (PlaScope, RFPlasmid, mlplasmids and Platon). The benchmark dataset consisted of 14,746 contigs. Of these contigs, 87.3% (n=12,872) were chromosome-derived and 12.7% (n=1,874) were plasmid-derived.

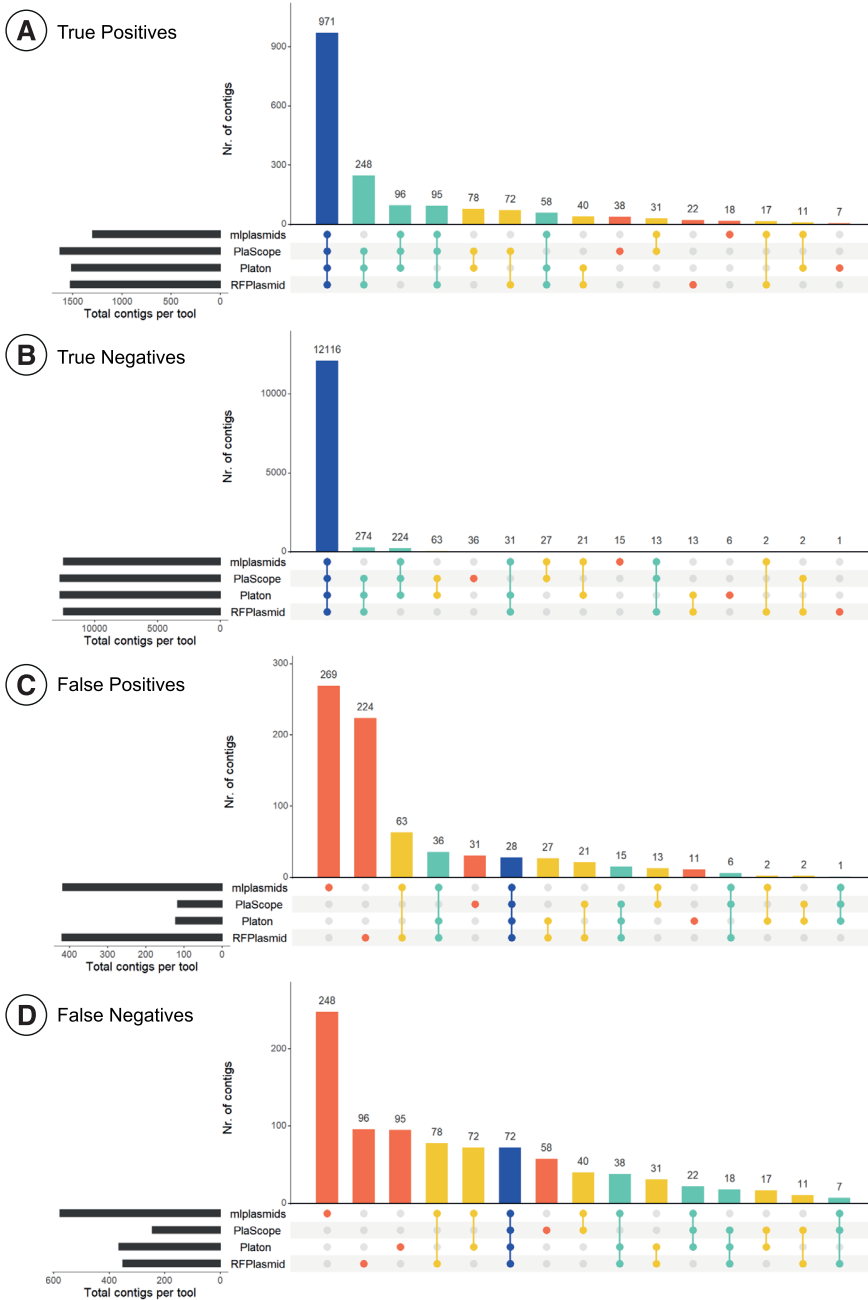
We evaluated the number of contigs correctly and incorrectly classified by each of the tools and calculated  $\text{recall}_{(\text{contig})}$ ,  $\text{precision}_{(\text{contig})}$  and  $\text{F1-score}_{(\text{contig})}$  (Supplementary Table S1). PlaScope was able to correctly identify the highest number of plasmid-derived contigs (True Positives, n=1,629), while the rest of the tools detected between 1,297 and 1,523 plasmid-derived contigs. Notably, PlaScope also included the least chromosomal contamination in its predictions (False Positives, 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 (n=418 and n=420, respectively). PlaScope was the tool with the highest  $\text{F1-score}_{(\text{contig})}$  (0.900) followed by Platon (0.861), RFPlasmids (0.798) and mlplasmids (0.722). For most tools,  $\text{precision}_{(\text{contig})}$  values were higher than  $\text{recall}_{(\text{contig})}$  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 (Figure 2). All tools agreed on the correct classification of 51.8% of plasmid-derived contigs (True Positives: n=971, Figure 2A), and another 26.5% plasmid-derived contigs were correctly classified by at least three tools (n=497). Also, a high fraction (94.1%) of chromosome-derived contigs were correctly classified by all tools (True Negatives: n = 12,116 contigs, Figure 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 (True Positives: 85/1,874, 4.7%, True Negatives: 58/12,872, 0.5% respectively). From these observations, we concluded that contigs misclassifications are primarily derived from individual tools (Figure 2C and 2D).

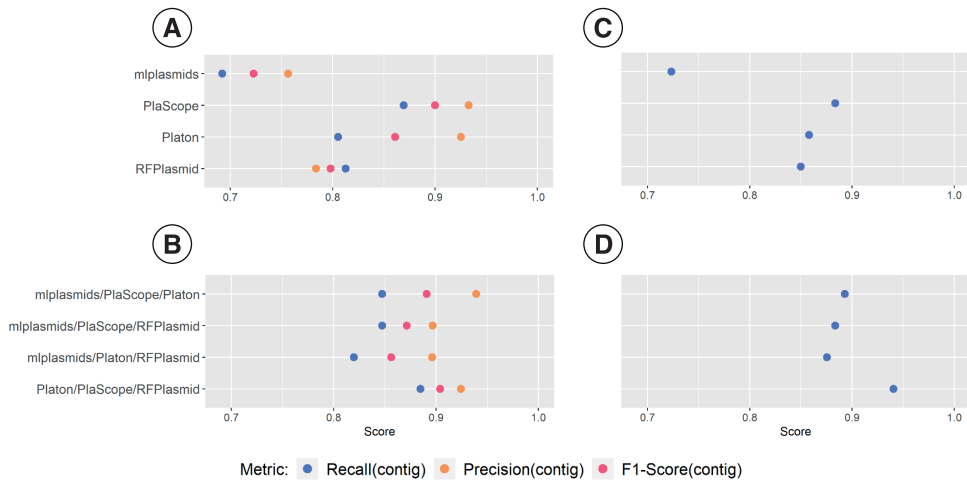
**PlasmidEC: A voting classifier for improved detection of ARG-plasmid contigs in *E. coli***

We theorised 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/PlaScope/RFPlasmid displayed the highest overall performance of voting classifiers with the highest  $\text{F1-score}_{(\text{contig})}$  (0.904). This ensemble classifier achieved an  $\text{F1-score}_{(\text{contig})}$  similar to PlaScope (0.900) but had a slightly higher  $\text{recall}_{(\text{contig})}$  (0.884 and 0.869, respectively) (Figure 3 A and B, Supplementary Table S1).



**Figure 2.** Upset diagrams showing congruence in contig classification by different binary prediction tools (absolute counts). Bar colours indicate the number of tools that are in congruence. Note that y-axes have different heights. True Positives (TP; prediction=plasmid, class=plasmid), True Negatives (TN; prediction = chromosome, class=chromosome), False Positives (FP; prediction=plasmid, class=chromosome), False Negatives (FN, prediction=chromosome, class=plasmid).



**Figure 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 full dataset (n=214 genomes). **B**) PlasmidEC combinations evaluated using full dataset. **C**) Individual classifiers evaluated for the identification of plasmid-derived contigs (recall<sub>(contig)</sub>) using dataset of ARG-plasmids (n=114 ARG-plasmids). **D**) PlasmidEC combinations evaluated for the identification of plasmid-derived contigs (recall<sub>(contig)</sub>) using dataset of ARG-plasmids.

Next, we evaluated recall<sub>(contig)</sub> values for a subset of plasmids (n=114) encoding antibiotic resistance genes (ARG-plasmids) (Figure 3C and 3D, Supplementary 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 (mplasmids) to 0.884 (PlaScope), whereas the different combinations of tools in a voting classifier reached recall<sub>(contig)</sub> values ranging from 0.883 (mplasmids/Platon/RFPlasmid) to 0.941 (Platon/PlaScope/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>.

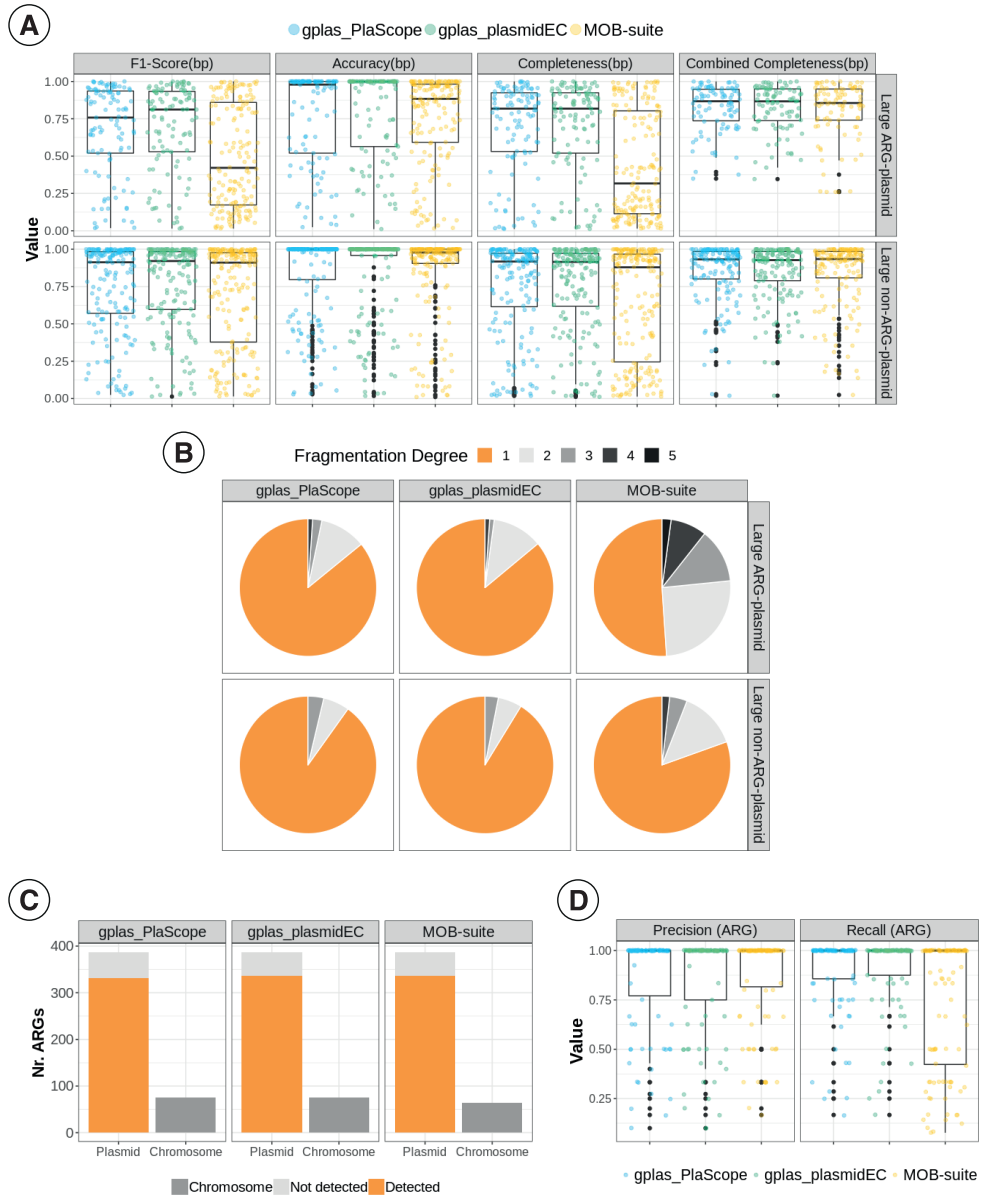
We measured the computational resources used by the ensemble and individual classifiers (Supplementary Figure S4). Binary classifiers showed considerable differences in both CPU time and memory used. The average CPU time required per sample was lowest for PlaScope (0.2 mins) and highest for Platon (14.9 mins). Platon also used the largest amount of memory per sample (20.6 Mb). The least amount of memory was required by mplasmids (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 mplasmids/PlaScope/RFPlasmid required the least number of resources (CPU time = 4.5 mins, memory = 9.0 Mb) and PlaScope/Platon/RFPlasmid the most (CPU time = 21.5 mins, memory = 21.4 Mb).

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

To reconstruct individual *E. coli* plasmids, gplas was combined with either plasmidEC or PlaScope, and performances were compared with MOB-suite, which was the best-performing plasmid reconstruction tool for *E. coli* in a recent benchmark study [19,34]. To retain comparability with the aforementioned study, we started with the same dataset and removed 26 genomes that were present in the PlaScope database and 15 genomes that were used to improve the gplas algorithm [35]. Consequently, our benchmark dataset consisted of 199 complete *E. coli* genomes encoding 483 plasmids. A total of 213 (44.1%) plasmids were classified as small plasmids (smaller than 18,000 bp), while the remaining 270 (55.9%) were large plasmids. 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  $F1\text{-Score}_{(bp)}$  values of gplas combined with either plasmidEC (gplas\_plasmidEC) or PlaScope (gplas\_PlaScope) were similar (Figure 4A, Table 1), with median  $F1\text{-Score}_{(bp)}$  of 0.81, (IQR=0.53 - 0.93) and 0.76 (IQR=0.52 - 0.94), respectively. Notably, both gplas methods outperformed MOB-suite, which presented a lower  $F1\text{-Score}_{(bp)}$  (median= 0.44, IQR= 0.18 - 0.87). As accuracy<sub>(bp)</sub> values were nearly identical across tools, the disparity in  $F1\text{-Scores}_{(bp)}$  can be explained due to the differences in completeness<sub>(bp)</sub>. In contrast, combined completeness<sub>(bp)</sub> distributions were virtually identical among tools. These results suggested that all methods had a similar capacity to detect contigs derived from large ARG-plasmids, but gplas 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 (Figure 4B). For large ARG plasmids, we found that MOB-suite fragmented 49% of plasmids into multiple predictions, while both gplas methods did so in only 14% of the cases.

All tools identified a similar number of plasmid-borne ARGs (Figure 4C). MOB-suite and gplas\_plasmidEC detected 331 (86.6%) ARGs encoded in large plasmids and gplas\_PlaScope 327 (85.6%). Moreover, all tools successfully detected all ARGs present in small plasmids ( $n=5$ , 100%). In concordance with previous results, recall<sub>(ARG)</sub> values (Figure 4D) for gplas predictions were higher than those obtained with MOB-suite (Table 1). This indicates that gplas performs better at correctly binning ARGs together into the same plasmid prediction. However, plasmid predictions made with gplas included a higher number of chromosomal ARGs (Figure 4C, Table 1).



**Figure 4.** Benchmarking of plasmid reconstruction methods for ARG-plasmids. **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 the contigs that compose a single 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	glas_plasmidEC	glas_PlaScope
<b>Large Plasmids (n = 270)</b>			
Nr. of detected plasmids*	263 (97.4%)	253 (93.7%)	254 (94.1%)
<b>ARG-Plasmids (n = 96)</b>			
F1-Score(bp) (median, IQR)	0.421 (0.172 - 0.860)	0.812 (0.529 - 0.934)	0.758 (0.520 - 0.936)
Completeness(bp) (median, IQR)	0.317 (0.114 - 0.803)	0.818 (0.520 - 0.924)	0.818 (0.531 - 0.924)
Accuracy(bp) (median, IQR)	0.883 (0.591 - 0.982)	0.979 (0.564 - 1)	0.979 (0.520 - 1)
Nr. plasmid-borne ARGs detected	331 (86.6%)	331 (86.6%)	327 (85.6%)
Nr. chromosome-derived ARGs	64	75	75
Recall (ARG) (median, IQR)	1 (0.42 - 1)	1 (0.86 - 1)	1 (0.86 - 1)
Precision (ARG) (median, IQR)	1 (0.82 - 1)	1 (0.75 - 1)	1 (0.77 - 1)
<b>Non-ARG-Plasmids (n = 174)</b>			
F1-Score(bp) (median, IQR)	0.910 (0.378 - 0.977)	0.921 (0.596 - 0.983)	0.912 (0.571 - 0.983)
Completeness(bp) (median, IQR)	0.879 (0.245 - 0.967)	0.915 (0.618 - 0.972)	0.918 (0.614 - 0.972)
Accuracy(bp) (median, IQR)	0.978 (0.904 - 1)	1 (0.958 - 1)	1 (0.796 - 1)
<b>Small Plasmids (n=213)</b>			
Nr. of detected plasmids*	174 (81.8%)	184 (86.4%)	196 (92.0%)
F1-Score(bp) (median, IQR)	1 (0.985 - 1)	1 (0.991 - 1)	1 (0.990 - 1)
Completeness(bp) (median, IQR)	1 (0.976 - 1)	1 (0.996 - 1)	1 (0.990 - 1)
Accuracy(bp) (median, IQR)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
Nr. plasmid-borne ARGs detected	5 (100%)	5 (100%)	5 (100%)

\*A plasmid is considered detected if at least 1 contig is included in the plasmid predictions.

Interestingly, tools performed comparably when evaluating the reconstruction of plasmids encoding extended spectrum beta-lactamases (ESBL) (n=42). MOB-suite reconstructions were characterised by having higher accuracy<sub>(bp)</sub> and gplas methods reconstructed ESBL-plasmids with higher completeness<sub>(bp)</sub> (Supplementary Figure S5A). Despite these differences, all tools exhibited similar F1-Score<sub>(bp)</sub> values. Additionally, the number of plasmid-borne ESBL genes detected were almost identical across tools (Supplementary Figure S5B). Nevertheless, gplas methods still performed better at binning ARGs into the same prediction (Supplementary Figure 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 (Table 1, Supplementary Figure S6A). This is likely due to most small plasmids being assembled into a single contig (n=196, 92.0%) (Supplementary Figure 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 (Supplementary Figure S6C). Interestingly, gplas\_PlaScope detected 196 (92.0%) small plasmids, and gplas\_plasmidEC performed similarly, detecting 184 (86.4%). Both gplas methods outperformed MOB-suite, which detected 174 (81.8%) small plasmids.

Finally, we tested the effect of using different contig sizes for plasmid reconstruction. We found no significant differences in performance of the tools when using 500 bp or 1,000 bp as a minimum contig size. A more detailed description of the results from this analysis can be found in Supplementary Materials and in Supplementary Figures S7 - S10.

## Discussion

In this work, we developed a new high-throughput method to reconstruct *E. coli* plasmids *de novo* from short-read sequencing data. Accurately reconstructing *E. coli* plasmids from Illumina reads has proven to be a challenge, especially in the context of ARG plasmids. 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 ( $\text{Recall}_{(\text{contig})}=0.941$ ), 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 for classifying contigs from ARG plasmids than from non-ARG plasmids (precision not analyzed), 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 PlaScope performed similarly in terms of  $\text{F1-Score}_{(\text{contig})}$ . However, plasmidEC showed a higher  $\text{Recall}_{(\text{contig})}$  but used more computational resources and took a longer time to complete. Reference-based methods, like PlaScope, 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 few novel plasmid sequences in a dataset that included more than 2,000 plasmids from *Enterobacteriaceae* isolates. PlaScope was built around Centrifuge [41], a metagenomic classifier to predict the origin of contigs based on custom databases. Recently, it was also shown that the usage of Kraken [42], another metagenomic classifier using customised databases, outperformed other binary classifiers in *Klebsiella pneumoniae* [41,43]. 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 optimised 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 relatively more to the result than a low confidence prediction.

We integrated plasmidEC (and PlaScope) with gplas to reconstruct individual *E. coli* plasmids. We then compared the performance of gplas combined with those classifiers against MOB-suite. Interestingly, the most pronounced differences in performance were observed when reconstructing ARG-plasmids. Although combined completeness<sub>(bp)</sub> values indicated that the three tools identified similar fractions of ARG-plasmids, MOB-suite more frequently fragmented ARG-plasmids into multiple bins, yielding low completeness<sub>(bp)</sub> and  $\text{F1-Score}_{(\text{bp})}$ . In contrast,



gplas (either with plasmidEC or PlaScope) was more successful at binning together contigs into individual plasmid predictions, thus achieving higher values for the used metrics. Accuracy<sub>(bp)</sub> values for all tools were similar, indicating a similar degree of chimeric predictions. Interestingly, both gplas 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 [18], 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 correct reconstruction of ARG-plasmids using coverage and node connectivity information. Yet, our study also proves that enriching the assembly graph with accurate information on the origin of nodes (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 gplas v1.0 could be combined with different binary classification tools to obtain accurate *de novo* plasmid reconstructions for multiple bacterial species. This means that gplas 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 in 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, gplas does not include repeated elements such as insertion sequences in the 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 [44], and they are frequently involved in the mobility of ARGs [9,45,46]. Additionally, the localization of these MGEs can influence the expression levels of ARGs [47,48], thereby impacting the resulting resistance phenotypes. Consequently, including IS elements would certainly improve the completeness and relevance of plasmid predictions. Some graph-based plasmid reconstruction methods like HyAsP [49], 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 [50], which refines binning results of metagenomics samples. A second disadvantage of our method is the formation of chimeras, which are bins composed of nodes from distinct replicons. As previously mentioned, accurate identification of plasmid derived nodes reduces the number of

chromosome-plasmid chimeras. However, preventing the formation of plasmid-plasmid chimeras is more challenging, especially for isolates carrying multiple large plasmids with similar copy numbers. Separating these chimeras could be possible with the use of a plasmid-backbone reference database.

To conclude, in this work we presented a new plasmidome prediction tool, named plasmidEC, and optimised gplas to accurately bin predicted plasmid sequences. Compared to existing binary classifiers, plasmidEC achieves increased recall, especially for contigs that derive from ARG plasmids. The integration of plasmidEC with gplas 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 values. The presented approach constitutes the best alternative to accurately predict and reconstruct ARG plasmids *de novo* in the absence of long-read data.

## References

1. 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.
2. Day MJ, Doumith M, Abernethy J, Hope R, Reynolds R, Wain J, 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.
3. Tumbarello M, Sanguinetti M, Montuori E, Trecarichi EM, Posteraro B, Fiori 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.
4. Mediavilla JR, Patrawalla A, Chen L, Chavda KD, Mathema B, Vinnard C, et al. Colistin- and Carbapenem-Resistant *Escherichia coli* Harboring *mcr-1* and *bla*NDM-5, Causing a Complicated Urinary Tract Infection in a Patient from the United States. *MBio.* 2016;7. doi:10.1128/mBio.01191-16
5. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet.* 2022. doi:10.1016/S0140-6736(21)02724-0
6. Jiang X, Ellabaan MMH, Charusanti P, Munck C, Blin K, Tong Y, et al. Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat Commun.* 2017;8: 15784.
7. Lermineaux NA, Cameron ADS. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can J Microbiol.* 2019;65: 34–44.
8. 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.
9. Che Y, Yang Y, Xu X, Brinda K, Polz MF, Hanage WP, 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. doi:10.1073/pnas.2008731118
10. Zhang S, Abbas M, Rehman MU, Huang Y, Zhou R, Gong S, et al. Dissemination of antibiotic resistance genes (ARGs) via integrons in *Escherichia coli*: A risk to human health. *Environ Pollut.* 2020;266: 115260.
11. 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.
12. Lopatkin AJ, Meredith HR, Srimani JK, Pfeiffer C, Durrett R, You L. Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat Commun.* 2017;8: 1689.
13. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. *Front Microbiol.* 2016;0. doi:10.3389/fmicb.2016.00173
14. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *Elife.* 2020;9. doi:10.7554/eLife.53886
15. Bosch T, Lutgens SPM, Hermans MHA, Wever PC, Schneeberger PM, Renders NHM, 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.
16. 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.
17. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun.* 2020;11. doi:10.1038/s41467-020-17278-2

18. 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.
19. 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.
20. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom.* 2018;4. doi:10.1099/mgen.0.000206
21. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom.* 2018;4. doi:10.1099/mgen.0.000211
22. 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. doi:10.1099/mgen.0.000683
23. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom.* 2020;6. doi:10.1099/mgen.0.000398
24. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom.* 2018;4. doi:10.1099/mgen.0.000224
25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol.* 2012;19: 455.
26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29: 1072–1075.
27. 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.
28. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother.* 2020;75: 3491–3500.
29. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. Available: <http://arxiv.org/abs/1303.3997>
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25: 2078.
31. Aaron R, Quinlan IMH. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26: 841.
32. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics.* 2020;36: 3874–3876.
33. Arredondo-Alonso S, Top J, McNally A, Puranen S, Pesonen M, Pensar J, et al. Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*. *MBio.* 2020;11. doi:10.1128/mBio.03284-19
34. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat Rev Microbiol.* 2021;19: 347–359.
35. Li Y, Luo Y. Performance-weighted-voting model: An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quantitative biology (Beijing, China).* 2020;8. doi:10.1007/s40484-020-0226-1
36. Millán AP, Alipour F, Hill KA, Kari L. DeLUCS: Deep learning for unsupervised clustering of DNA sequences. *PLoS One.* 2022;17. doi:10.1371/journal.pone.0261531
37. Wattanapornprom W, Thammarongtham C, Hongsthong A, Lertampaiporn S. Ensemble of Multiple Classifiers for Multilabel Classification of Plant Protein Subcellular Localization. *Life.* 2021;11. doi:10.3390/life11040293

38. Xue T, Zhang S, Qiao H. i6mA-VC: A Multi-Classifer Voting Method for the Computational Identification of DNA N6-methyladenine Sites. *Interdiscip Sci.* 2021;13. doi:10.1007/s12539-021-00429-4
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;0. doi:10.3389/fmicb.2020.00483
40. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated *Enterobacteriaceae*. *Sci Adv.* 2021;7. doi:10.1126/sciadv.abe3868
41. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 2016. doi:10.1101/gr.210641.116
42. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. doi:10.1101/762302
43. Gomi R, Wyres KL, Holt KE. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microbial genomics.* 2021;7. doi:10.1099/mgen.0.000550
44. 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.
45. Razavi M, Kristiansson E, Flach C-F, Joakim Larsson DG. The Association between Insertion Sequences and Antibiotic Resistance Genes. *mSphere.* 2020;5. doi:10.1128/mSphere.00418-20
46. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev.* 2018;31. doi:10.1128/CMR.00088-17
47. Kamruzzaman M, Patterson JD, Shoma S, Ginn AN, Partridge SR, Iredell JR. Relative Strengths of Promoters Provided by Common Mobile Genetic Elements Associated with Resistance Gene Expression in Gram-Negative Bacteria. *Antimicrob Agents Chemother.* 2015;59. doi:10.1128/AAC.00420-15
48. Turton JF, Ward ME, Woodford N, Kaufmann ME, Pike R, Livermore DM, et al. The role of ISAbal in expression of OXA carbapenemase genes in *Acinetobacter baumannii*. *FEMS Microbiol Lett.* 2006;258. doi:10.1111/j.1574-6968.2006.00195.x
49. Müller R, Chauve C. HyAsP, a greedy tool for plasmids identification. *Bioinformatics.* 2019;35: 4436–4439.
50. Mallawaarachchi VG, Wickramarachchi AS, Lin Y. Improving metagenomic binning results with overlapped bins using assembly graphs. *Algorithms Mol Biol.* 2021;16: 3.

## Supplementary Materials

### Supplementary Data

Supplementary data be downloaded from: <https://doi.org/10.5281/zenodo.7926472>

### Supplementary Results

#### Fractions of large plasmids can be found on nodes smaller than 1kb

We used nodes larger than 1kb as an input for gplas and MOB-suite. When aligning all these nodes to their corresponding complete genomes, we discovered that a considerable fraction of certain plasmids was missing (Supplementary Figure S7). The recovered fraction for small plasmids (median=1, IQR=0.98 - 1) was generally higher than for large plasmids (median=0.96, IQR= 0.91 - 0.99).

After including nodes with sizes ranging from 500 bp to 1 kb, we observed an increase in the recovered fraction to a total of 201 plasmids. This increase mainly occurs in large plasmids (n=194, 71.9%) and rarely in small plasmids (n=7, 3.3%) (Supplementary Figure S7). However, the relative increase in recovered fraction within large plasmids was minimal (median=0.98, IQR=0.94 - 1).

Despite the inclusion of smaller contigs, the recovered fraction remained below 0.9 for 26 (9.6%) large plasmids. Additionally, a total of 2 small plasmids and 1 large plasmid were entirely missing after assembly (recovered fraction = 0).

In order to determine if the missing plasmids or plasmid regions were successfully sequenced, we aligned Illumina reads back to their complete genomes and analyzed the sequencing coverage distribution. For simplification, we show results for plasmids that had a recovered fraction below 0.8 (n=11). Most of these isolates contained multiple plasmids, presenting an overall median of 6 plasmids. We found Illumina reads aligning to all bases that were missed from assembly (Supplementary Figure S8). Interestingly, in 5 of these plasmids (CP051632.1, CP055630.1, AP022225.1, AP022249.1, CP054283.1) the median sequencing coverage was lower than the median coverage of the chromosome. In the remaining 6 cases, however, there was no apparent correlation between the coverage in unassembled regions and the chromosome median coverage (Supplementary Figure S9).

Given that recovered fractions for large plasmids increased when including smaller contigs, we re-run all plasmid prediction tools including these contigs as input. In contrast to expectation, the overall performance of the plasmid reconstruction tools did not improve and remained almost identical for every metric (Supplementary Figure S10).

## Supplementary Tables

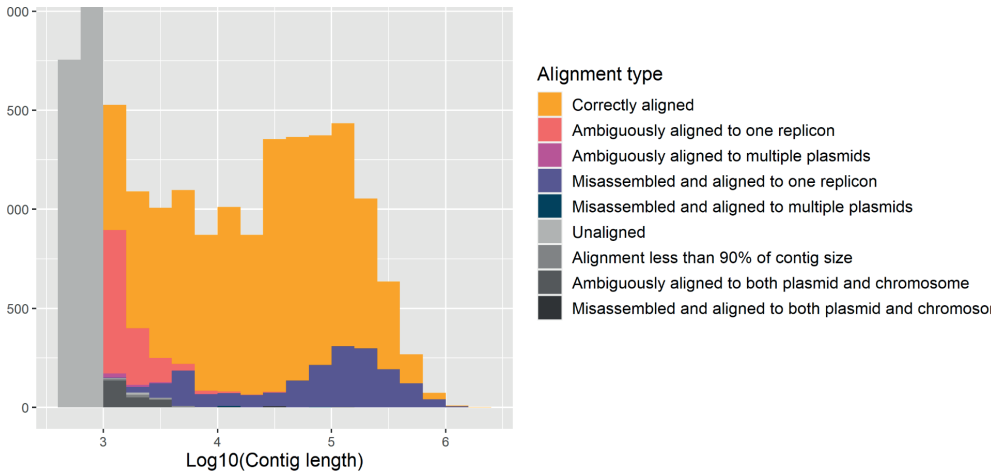
**Supplementary Table S1.** Performance metrics of binary classifiers and plasmidEC combinations evaluated for complete benchmarking dataset. Predictions were categorised into: True Positives (TP, prediction = plasmid, class = plasmid), True Negatives (TN, prediction = chromosome, class = chromosome), False Positives (FP, prediction = plasmid, class = chromosome) and False Negatives (FN, prediction = chromosome, class = plasmid).

Software	TP	TN	FP	FN	Precision	Recall	F1-Score
mplasmids	1,297	12,449	418	577	0.756	0.692	0.722
PlaScope	1,629	12,755	117	245	0.932	0.869	0.900
Platon	1,509	12,748	122	365	0.925	0.805	0.861
RFPlasmid	1,523	12,452	420	351	0.783	0.812	0.798
mplasmids/PlaScope/RFPlasmid	1,588	12,689	183	286	0.896	0.847	0.871
mplasmids/Platon/PlaScope	1,588	12,769	103	286	0.939	0.847	0.890
mplasmids/PlatonRFPlasmid	1,536	12,694	178	338	0.896	0.819	0.560
Platon/PlaScope/RFPlasmid	1,658	12,736	136	216	0.924	0.884	0.904

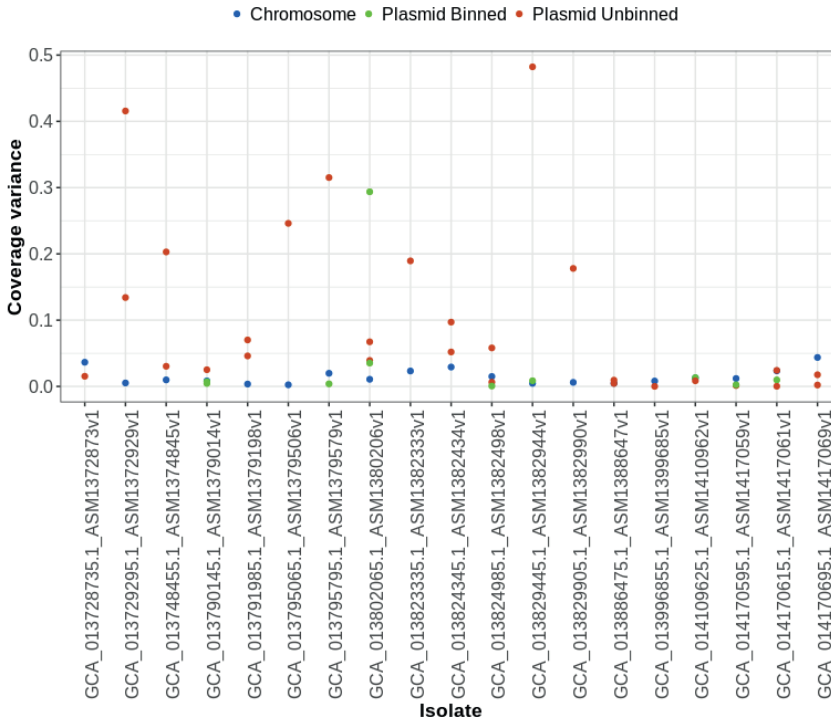
**Supplementary Table S2.** Performance metrics of binary classifiers and plasmidEC combinations evaluated for ARG-plasmids. Predictions were categorised into: True Positives (TP, prediction = plasmid, class = plasmid) and False Negatives (FN, prediction = chromosome, class = plasmid).

Software	TP	FN	Recall
mplasmids	622	238	0.723
PlaScope	760	100	0.883
Platon	738	122	0.858
RFPlasmid	731	129	0.80
mplasmids/PlaScope/RFPlasmid	760	100	0.883
mplasmids/Platon/PlaScope	768	92	0.893
mplasmids/PlatonRFPlasmid	753	107	0.875
Platon/PlaScope/RFPlasmid	809	51	0.941

Supplementary Figures

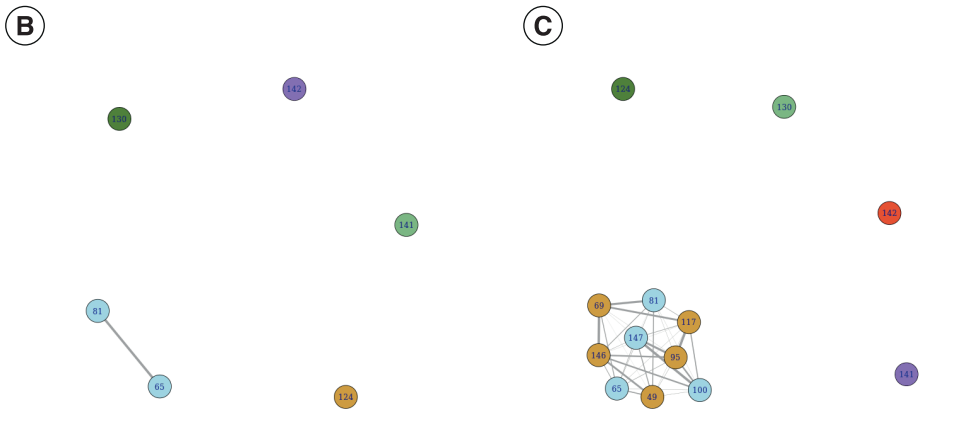
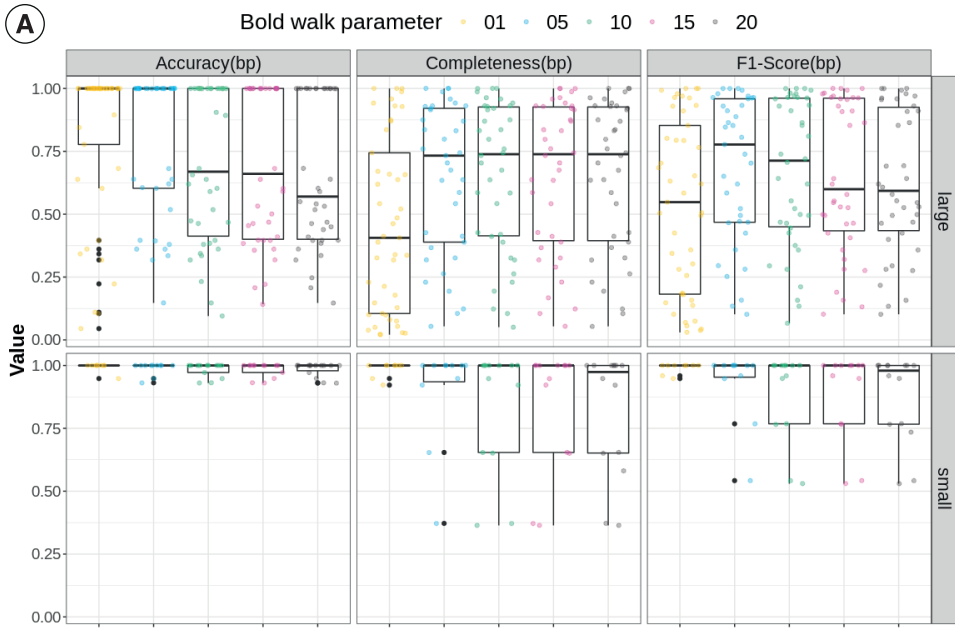


Supplementary Figure S1. Alignment types of all contigs in the dataset by contig length. Included contigs are shown in colour, excluded contigs in greyscale.

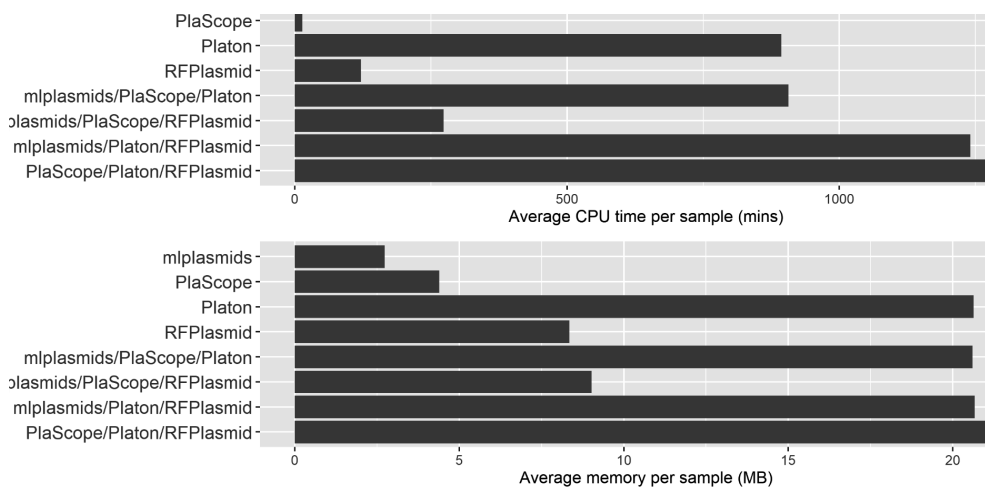


Supplementary Figure S2. Contig coverage variance for all replicons carried by isolates that contained unbinned nodes after gplas\_plasmidEC prediction.

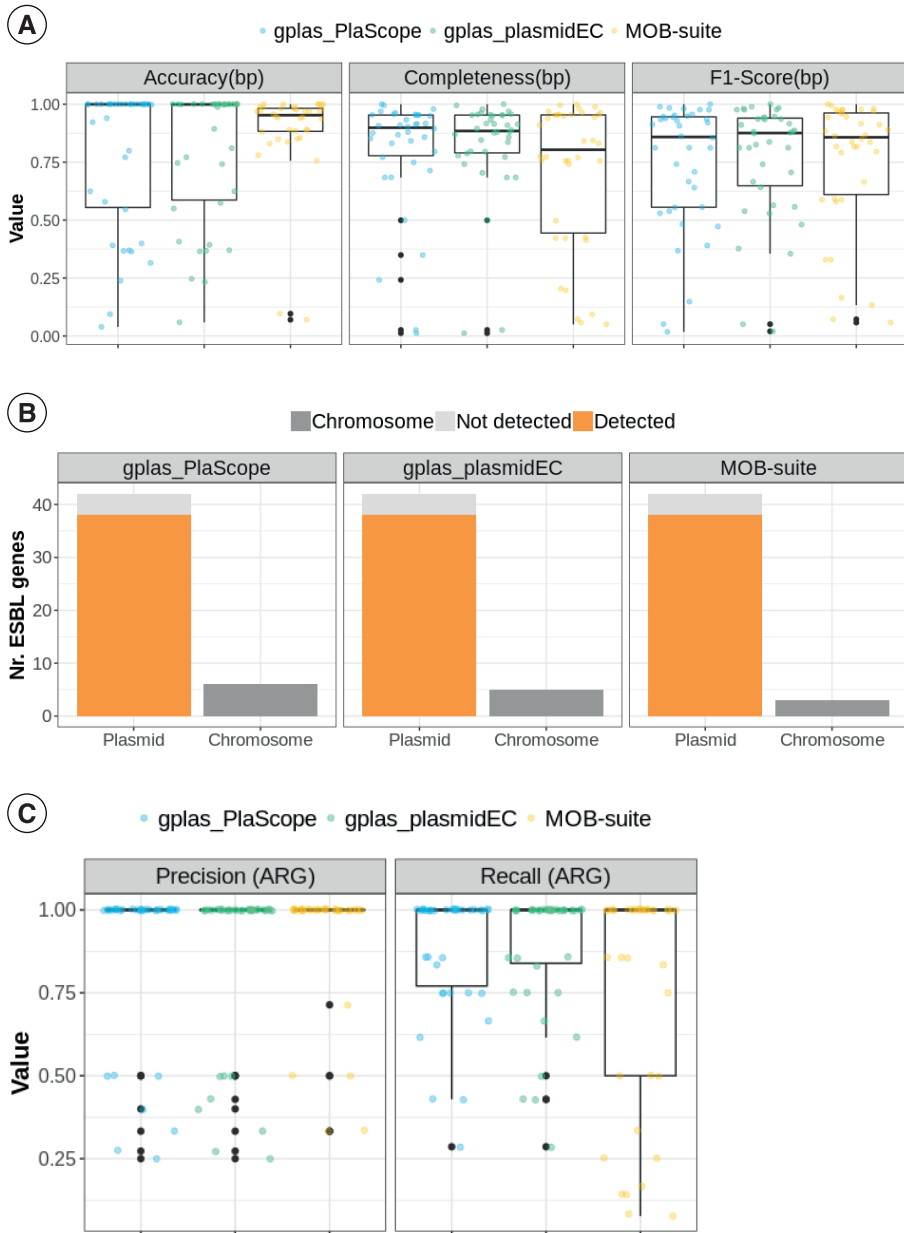




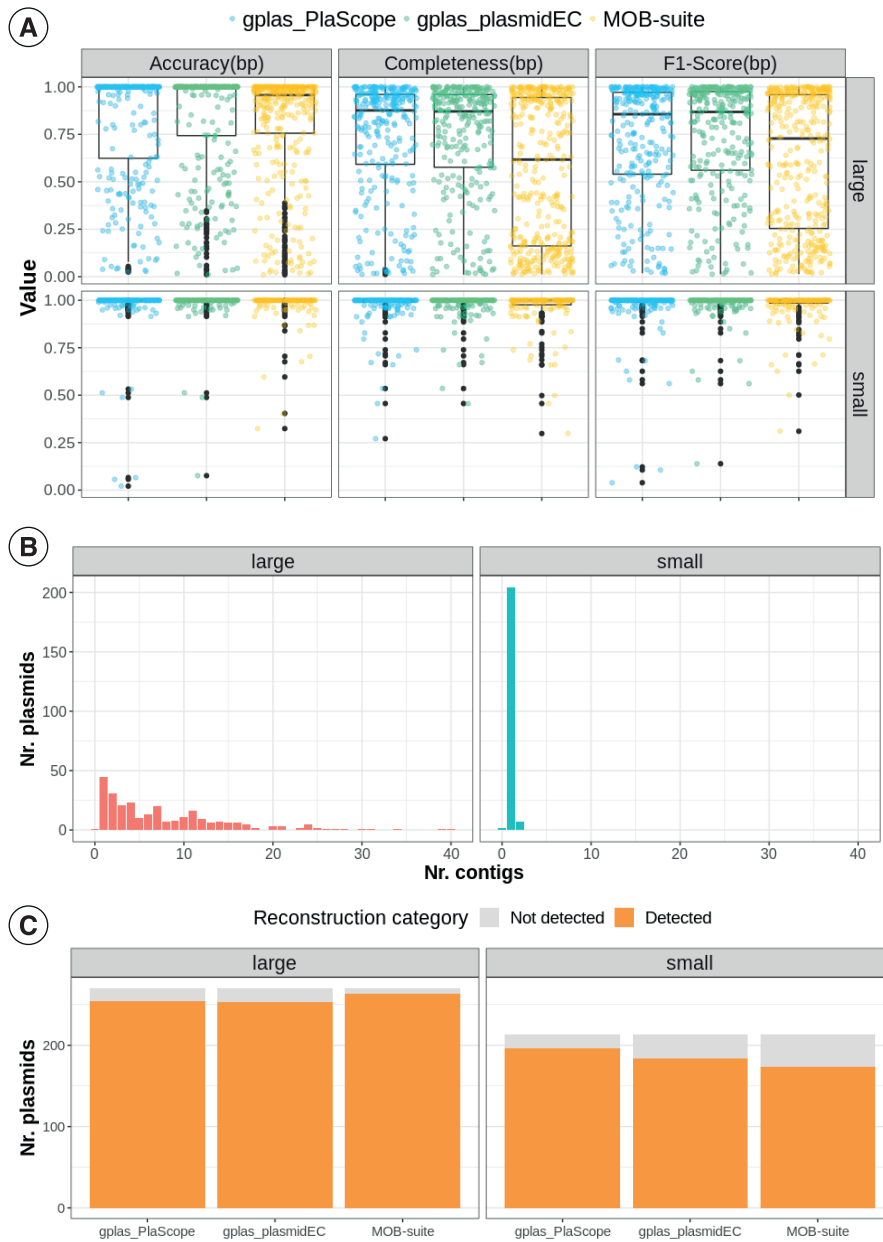
**Supplementary Figure S3.** A) Completeness(bp), Accuracy(bp) and F1-score(bp) values for plasmid predictions derived from isolates in which unbinned unitigs were predicted by gplas\_plasmidEC (n=15). gplas\_plasmidEC was run allowing different coverage variances in the bold mode. B) Plasmidome network obtained after running gplas without the bold parameter on isolate GCA\_013823335.1\_ASM1382333v1. Circles represent unitigs predicted as plasmid by plasmidEC and binned by gplas. Different colours correspond to different individual predicted plasmids. C) Plasmidome network obtained after running gplas with bold parameter of 5, on the same isolate.



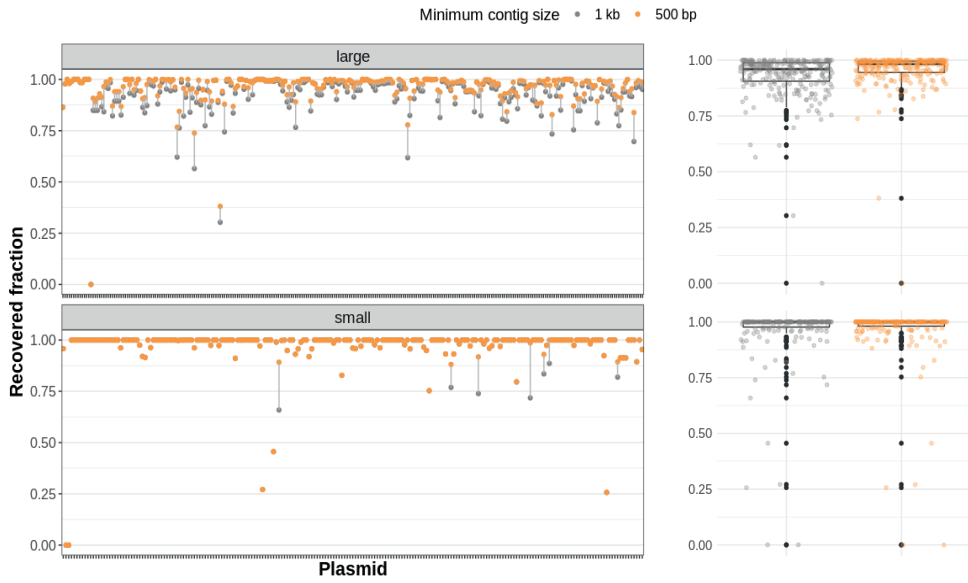
**Supplementary Figure S4.** Average computational resources used per sample: CPU time in minutes (A) and memory in Mb (B). Softwares were run on the full benchmarking dataset (n = 214).



**Supplementary Figure S5.** A) Completeness(bp), Accuracy(bp), F1-score(bp) values for predictions carrying plasmid-derived ESBL genes (n=42). B) Absolute count of plasmid-derived ESBL genes included (detected) and missed (Not detected) in plasmid predictions. Absolute count of chromosome-derived ESBLs contaminating plasmid predictions are also depicted. C) Recall(ARG) and Precision(ARG) values for ESBL-carrying plasmids.



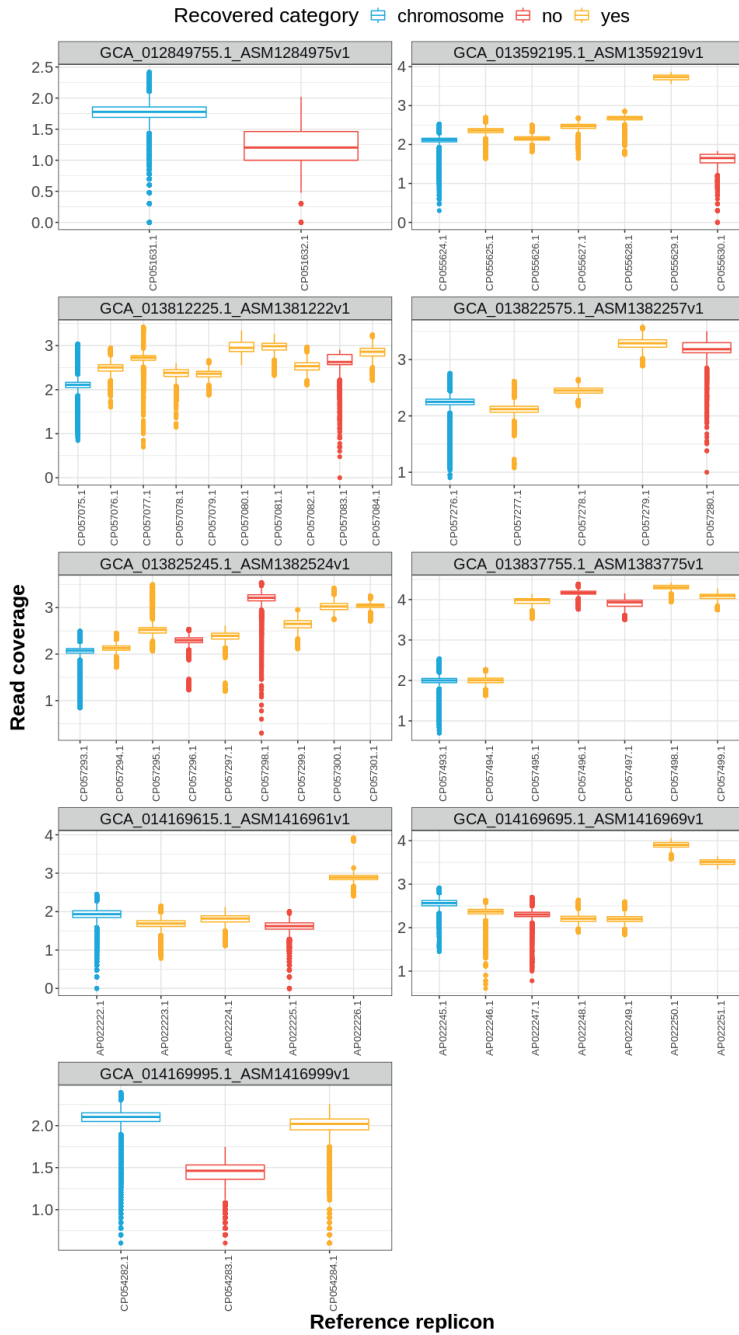
**Supplementary Figure S6.** A) Completeness(bp), Accuracy(bp), F1-score(bp) values for plasmid predictions according to size category (definition: small plasmids < 18 kb, large plasmids ≥ 18 kb). B) Histogram showing the number of contigs that compose each reference plasmid when these are assembled from short reads only. C) Absolute count of plasmids detected and undetected by each of the tools according to plasmid size. A reference plasmid was labelled as detected when at least one of its contigs was included into the predictions.



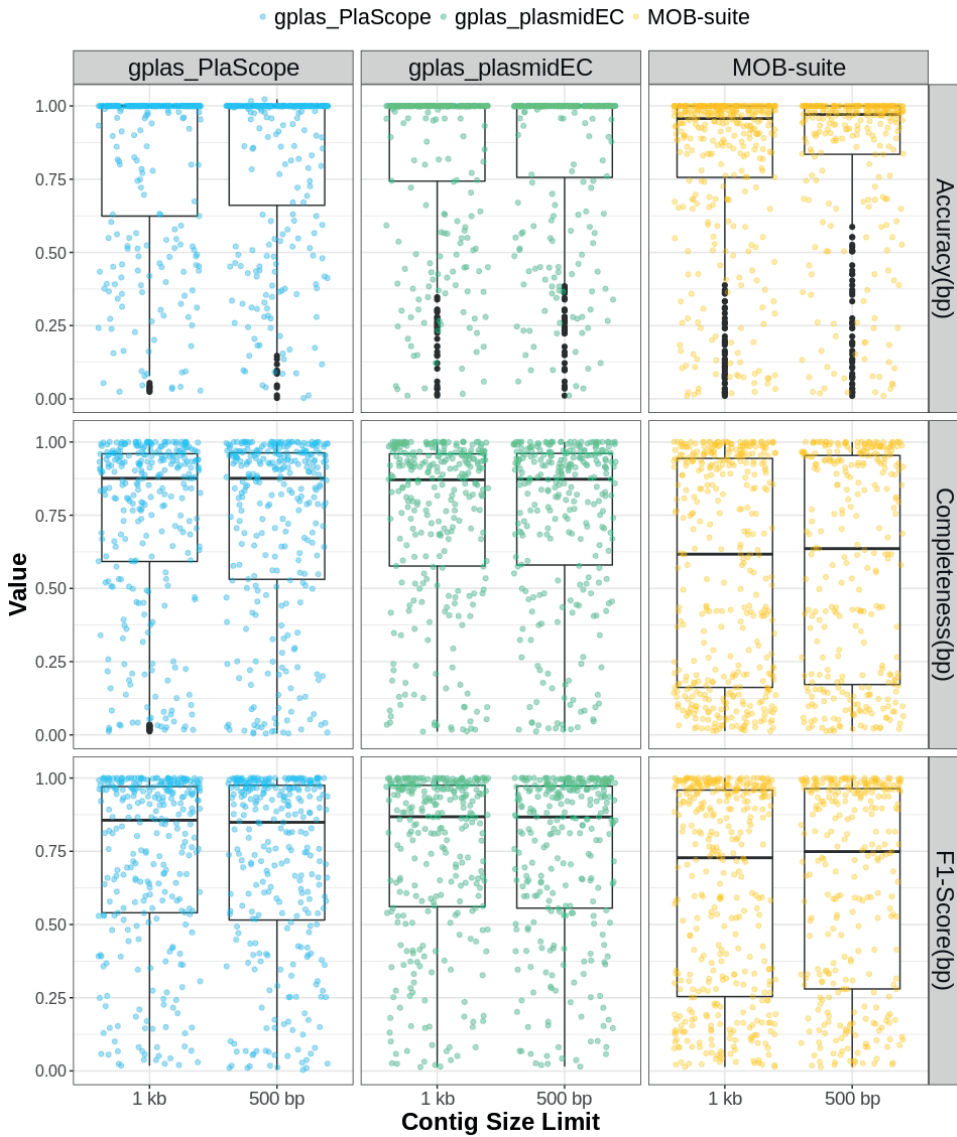
**Supplementary Figure S7.** Recovered fraction of plasmids after aligning all contigs larger than 500 bp (orange) or larger than 1 kb (grey) to the complete genomes.



Supplementary Figure S8. Read coverage per replicon position for plasmids that had a recovered fraction equal or smaller than 0.8. Regions that were not assembled in contigs larger than 500 bp are shown in red. Blue dotted line represents the median coverage for the chromosome.



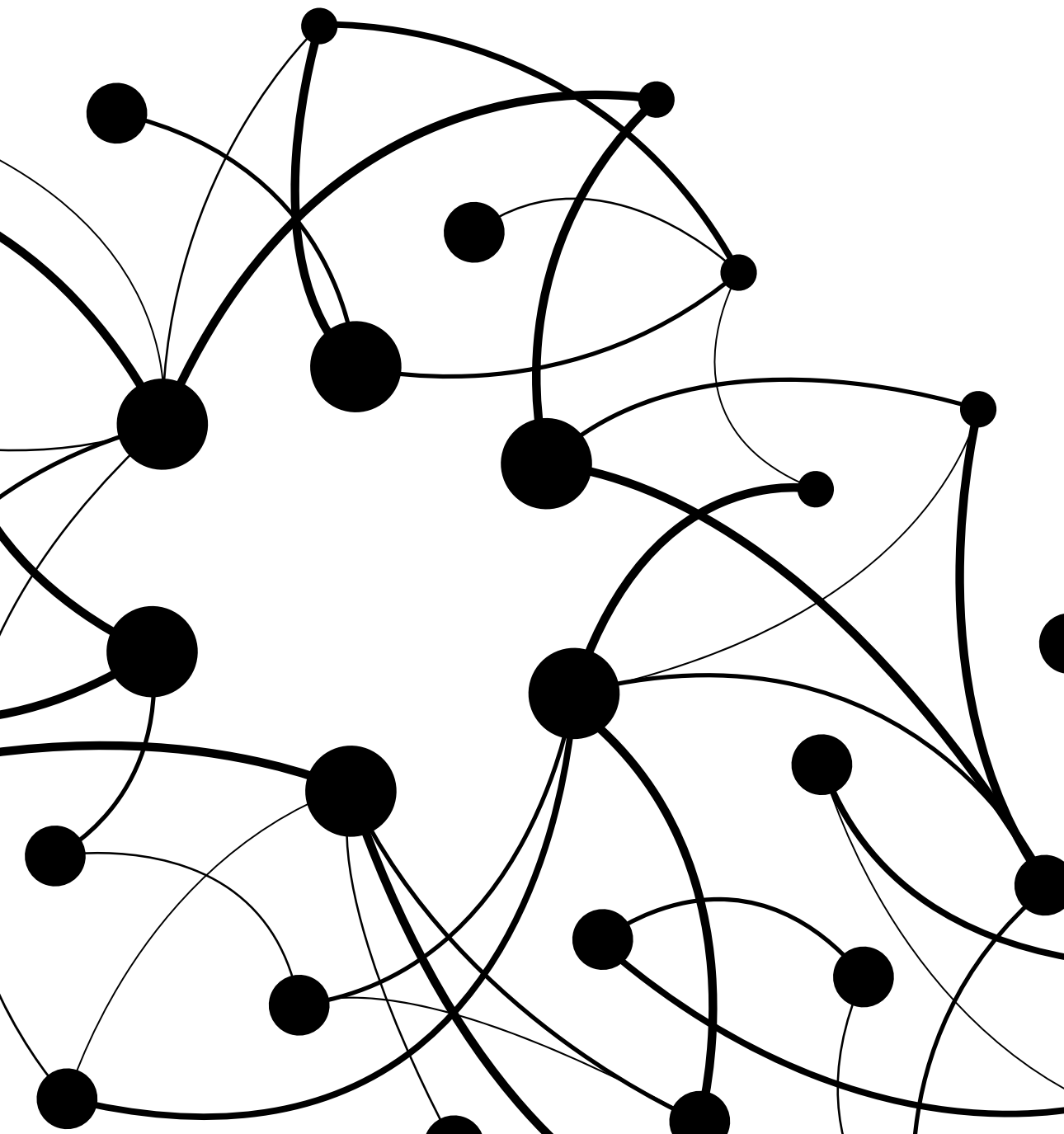
Supplementary Figure S9. Read coverage distribution for all replicons carried by isolates that had at least one plasmid with a recovered fraction <0.8 (n=11).



**Supplementary Figure S10.** Distribution of Accuracy(bp), Completeness(bp) and F1-Score(bp) values for plasmid predictions obtained using as input contigs of sizes larger than 500 bp or larger than 1 kb. Only large plasmids (n=270) were included in this analysis.







# 04

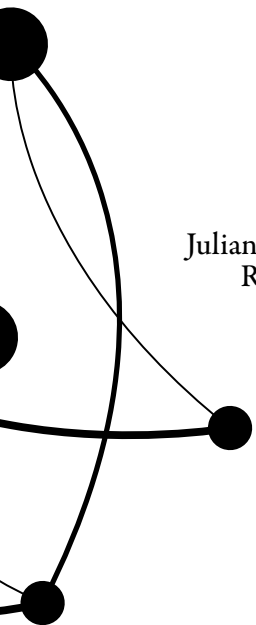
---

## **Accurately reconstructing AMR plasmids of multiple bacterial species from short reads**

Julian A. Paganini\*, Jesse Kerkvliet\*, Nienke L. Plantinga, Rodrigo Meneses,  
Rob J.L. Willems, Sergio Arredondo-Alonso and Anita C. Schürch

\*Authors contributed equally

Manuscript in preparation



### Abstract

Plasmids play a pivotal role in the spread of antibiotic resistance genes. Accurately reconstructing plasmids often requires long-read sequencing, which is more expensive and currently still more error-prone than short-read sequencing. We recently presented an optimised approach for reconstructing *Escherichia coli* antimicrobial resistance plasmids using Illumina short reads. This method consists of combining a robust binary classification tool named plasmidEC, with gplas, which is a tool that makes use of features of the assembly graph to bin predicted plasmid contigs into individual plasmids. Here, we developed different plasmidEC models: four species-specific models (*Enterococcus faecium*, *Klebsiella pneumoniae*, *Staphylococcus aureus* and *Salmonella enterica*) and one species-independent model for less frequent species. We combined these models with gplas to reconstruct plasmids from more than 70 different bacterial species.

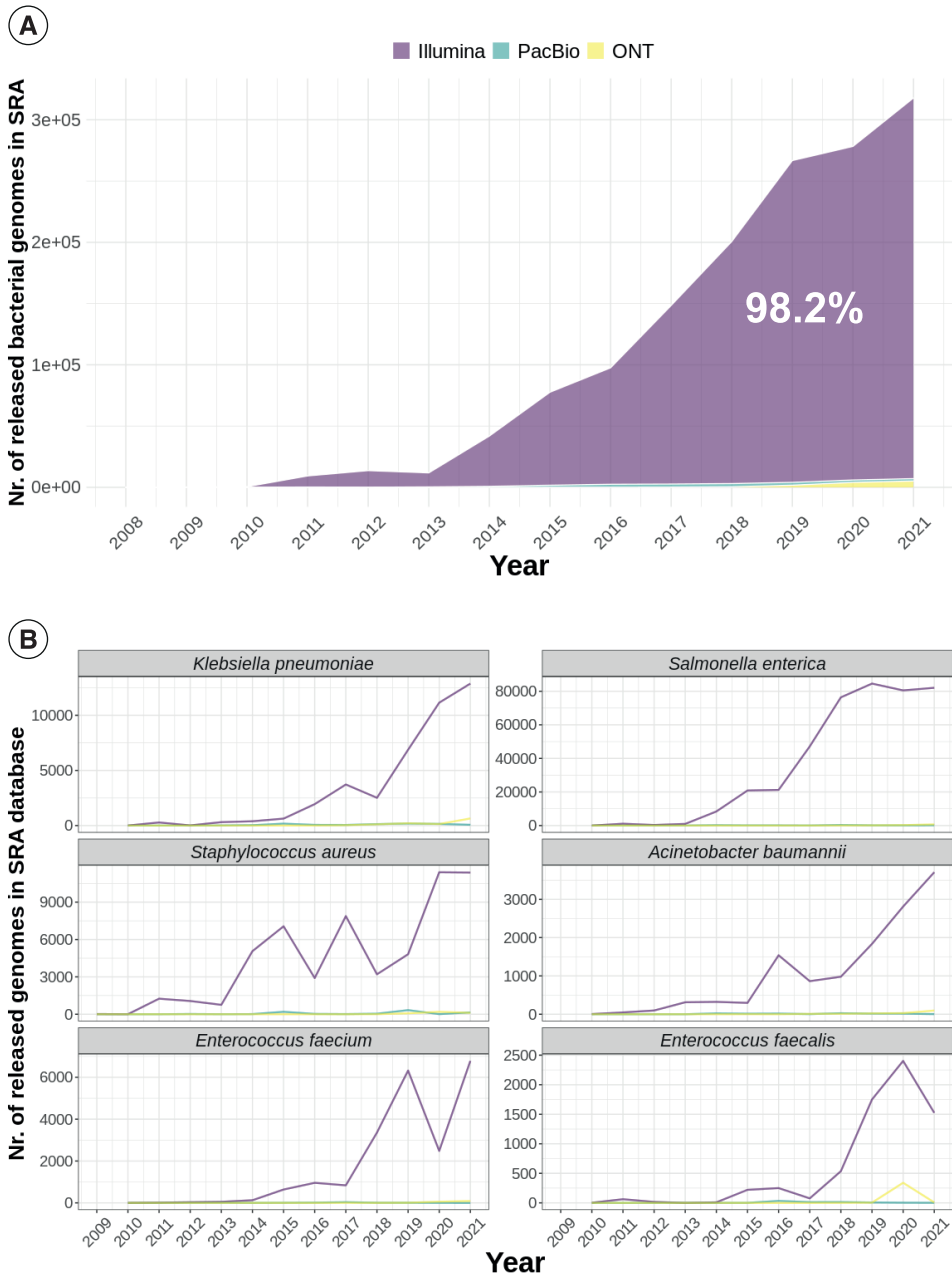
## Introduction

Antimicrobial resistance (AMR) is a major threat to human health. Recent estimates indicate that 1.27 million deaths were attributable to bacterial AMR in 2019 alone. Moreover, the number of infections caused by resistant bacteria is increasing each year [1]. For these reasons, bacterial AMR is now called the ‘silent pandemic’. Only a small number of new antibiotics have been approved by the FDA in recent years [2], and their use is recommended only for a limited number of clinical scenarios [3]. Although alternative approaches to treat bacterial infections are being explored, their effectiveness has not been extensively tested to date [4–7], and it will take several years until these methods become commonplace to treat bacterial infections. Given this scenario, limiting the dissemination of resistance is the key to preventing an AMR crisis.

The spread of AMR is a complex phenomenon that depends on various factors. However, it is known that plasmids play a central role in this process. Plasmids are mobile genetic elements (MGE) that frequently carry resistance genes and that can be transferred between bacteria by diverse mechanisms [8–12]. Several studies have described that plasmids also play an important role in the development of outbreaks in clinical settings that involve multiple bacterial species [13–17]. Therefore, accurate, high-throughput plasmid identification and tracking are becoming increasingly necessary.

Next-generation sequencing (NGS) platforms offer powerful tools for large-scale bacterial genomes research. Despite the recent advent of long-reads technologies, which allow obtaining complete bacterial genomes [18,19], Illumina short reads remains the most widespread sequencing method. As of July 2022, the Sequence Read Archive (SRA) contained more than 1.8 million DNA sequences belonging to bacterial isolates (Supplementary Data 1) and 98.2% of these were obtained using short-read technology (Figure 1A). Nevertheless, plasmids commonly contain repeat elements, which complicates their assembly with the use of short-read data alone [20]. Therefore, new and improved methods for reconstructing plasmids using short reads alone are needed.

Recently, we have developed a new method for reconstructing individual *E. coli* plasmids with short reads (Chapter 3). Briefly, nodes in the assembly graph are initially classified as plasmid- or chromosome-derived using plasmidEC, an ensemble classifier that combines the output from three existing binary classification tools [21–23]. Second, we used gplas [24] to bin plasmid nodes into individual plasmid predictions based on similarities in sequence coverage and graph connectivity. Our method performed better than MOB-suite [25], especially when reconstructing antibiotic-resistance gene (ARG) carrying plasmids.



**Figure 1.** A) Number of genomes sequenced by different NGS technologies, publicly available in the Sequence Read Archive (SRA) by 26 July 2022. 98.2% of genomes were sequenced with Illumina short-read technology, while 1% of sequences were obtained with Oxford Nanopore Technology (ONT) and 0.8% with Pacific Biosystems (PacBio). B) Same as A, but for individual species.

In this work, we aimed at improving the reconstruction of plasmids in multiple bacterial species using short reads. In order to achieve this, we followed the steps detailed below:

- 1) We developed four species-specific plasmidEC models to obtain binary classifications (plasmid/chromosome) of contigs in common human pathogens [1,26,27] that are highly represented in databases (*Enterococcus faecium*, *Klebsiella pneumoniae*, *Salmonella enterica* and *Staphylococcus aureus*) (Figure 1B). Additionally, we designed one species-independent model to classify contigs of all remaining bacterial species. We benchmarked these models against other binary classification tools.
- 2) We combined the best binary classification tool for each of these five models with gplas in order to reconstruct individual plasmids (n=953) from more than 70 species, focusing mainly on ARG-plasmid reconstructions. Predictions generated with gplas were compared against those of MOB-suite and plasmidSPAdes [25,31].

Finally, we also evaluated the impact of different plasmidome characteristics on plasmid reconstruction capabilities.

## Methods

All scripts used to reproduce the analyses can be found at [gitlab.com/jpaganini/reconstructing\\_amr\\_plasmids](https://gitlab.com/jpaganini/reconstructing_amr_plasmids). R version 3.6.1. was used for all R scripts.

### Development of plasmidEC models

PlasmidEC is an ensemble of three binary classification tools that implements a majority vote system to predict the origin of contigs (plasmid/chromosome).

For *E. faecium*, the plasmidEC model combines the outputs from RFPlasmid (v0.0.18) [23], Platon (v1.6) [21] and mlplasmids (v2.1.0) [21]. For *K. pneumoniae*, *S. aureus*, *S. enterica* and for the species-independent model, plasmidEC combines RFPlasmid, Platon and newly developed Centrifuge-based classifiers [30].

### Building Centrifuge-based classifiers

Centrifuge [30] is a tool that serves as a taxonomy classifier for metagenomics reads. We adapted this tool to function as a binary classifier of WGS bacterial contigs by building custom databases with complete bacterial sequences labelled as plasmid or chromosome, similarly to what has been described for PlaScope [22]. In contrast to PlaScope, our Centrifuge-based classifier calculates the proportion of hits from each contig in relation to different fractions of the database in order to generate the classification. For example, if a contig matches more than 70% of the times to the plasmidome fraction of the database, it is classified as a 'plasmid' contig. In contrast, if it matches less than 30%, it is classified as a 'chromosome' contig. Matches ranging from 30 to 70% are labelled 'unclassified'.

We built three species-specific databases for classifying contigs of *K. pneumoniae*, *S. aureus* and *S. enterica*. These databases were built using complete genomes uploaded to RefSeq prior to May 2020 (N=2,532 genomes). The number of genomes included from each species can be found in Supplementary Figure S1A, and the corresponding metadata is reported in Supplementary Data 2. A spe-

cies-independent (General) database was also built, using all available genomes from any bacterial species uploaded to RefSeq prior to 2022 and genomes uploaded in 2022 that did not have their corresponding short reads in the SRA database. A total of 25,735 genomes were included in this model. A summary of the included species can be found Supplementary Figure S1B. All databases are publicly available at zenodo.org and can be downloaded from the following links:

- [https://zenodo.org/record/7194565/files/K\\_pneumoniae\\_plasmid\\_db.tar.gz](https://zenodo.org/record/7194565/files/K_pneumoniae_plasmid_db.tar.gz)
- [https://zenodo.org/record/7133407/files/S\\_enterica\\_plasmid\\_db.tar.gz](https://zenodo.org/record/7133407/files/S_enterica_plasmid_db.tar.gz)
- [https://zenodo.org/record/7133406/files/S\\_aureus\\_plasmid\\_db.tar.gz](https://zenodo.org/record/7133406/files/S_aureus_plasmid_db.tar.gz)
- [https://zenodo.org/record/7431957/files/general\\_plasmid\\_db.tar.gz](https://zenodo.org/record/7431957/files/general_plasmid_db.tar.gz)

### **Benchmarking the performances of plasmidEC and gplas**

#### Compiling benchmarking datasets

PlasmidEC was benchmarked against RFPlasmid, Platon, Centrifuge and mlplasmids, while gplas performance was compared to that of MOB-suite and plasmidSPAdes.

To validate the performance of the tools, we compiled a benchmark dataset that consisted of 809 complete bacterial genomes with 1,923 plasmids. Of these, 492 genomes and 1,314 plasmids belonged to four common human pathogens that are highly abundant in sequence databases (*E. faecium*, *K. pneumoniae*, *S. enterica* and *S. aureus*) (Supplementary Figure S2A). The remaining 317 complete genomes and 609 plasmids belonged to less common species (Supplementary Figure S2B). The benchmark dataset was randomly divided into two groups containing an equal number of genomes per species (Supplementary Figure S3). Benchmark dataset “A” was used to compare the performance of plasmidEC to other binary classifiers. Benchmark dataset “B” was used to compare the performance of gplas to MOB-suite and plasmidSPAdes. A detailed list of genomes included in each benchmark dataset can be found in Supplementary Data 3.

To exclude genomes that were used in the development of MOB-suite, we included in our benchmark dataset only those genomes from common species that were uploaded to RefSeq after May 2020. For less frequent species, we included genomes uploaded in 2022.

Plasmids from *E. faecium*, *K. pneumoniae*, *S. enterica* and *S. aureus* captured most of the available plasmidome diversity for each species (Supplementary Figure S4).

Complete genomes and corresponding short reads were downloaded from RefSeq and SRA using ncbi-genome-download (v0.2.10) (<https://github.com/kbclin/ncbi-genome-download>) and SRA tools (v2.10.9), respectively.

#### Exploring the plasmid diversity of common species in the benchmarking datasets

We used Mash v2.2.2 ( $k = 21$ ,  $s = 10,000$ ) [34] to estimate the pairwise  $k$ -mer distances of all complete plasmid sequences of *K. pneumoniae*, *E. faecium*, *S. aureus* and *S. enterica* deposited in RefSeq. The obtained distances were clustered using the  $t$ -distributed stochastic neighbour embedding algorithm ( $t$ -SNE) with a perplexity value of 30. Data points, which represent individual sequences of plasmids, were coloured if they were part of the benchmarking dataset.



### Genome assembly and quality trimming

Illumina raw reads corresponding to the benchmark dataset were trimmed using trim-galore (v0.6.6) (<https://github.com/FelixKrueger/TrimGalore>) to remove adapter contamination and bases with a phred quality score below 20. Unicycler (v0.4.8) [35] was then applied to perform *de novo* assembly with default parameters.

### Determining the origin of all assembled contigs

After assembly with Unicycler, the resulting contigs were labelled as chromosome- or plasmid-derived by alignment to their corresponding complete genomes using QUAST (v5.0.2)[36]. Only contigs larger than 1,000 bp with an alignment of at least 90% the contig length were considered (n=59,380). Of those, contigs aligning to both the chromosome and plasmidome (ambiguously aligned contigs) were discarded (n=1,187) In total, excluded contigs represented 2.8% of the entire dataset.

### Evaluating binary classification tools

We evaluated the performance of all binary classifiers by comparing, for each contig, their prediction with the actual class of the contig. For Centrifuge, all 'unclassified' predictions were considered chromosomes. The predictions were categorised as follows: 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:

$$Recall (contig) = \frac{TP}{TP + FN}$$

$$Precision (contig) = \frac{TP}{TP + FP}$$

$$F1 - Score (contig) = \frac{2 \times Recall (contig) \times Precision (contig)}{Recall (contig) + Precision (contig)}$$

### Evaluating plasmid reconstruction tools

To reconstruct individual plasmids, gplas was combined with different binary classification tools, based on their performance. To recover plasmids from *K. pneumoniae* and less-frequent species, Centrifuge was selected, while for the other three species, plasmidEC (v1.3) was chosen, using the majority vote.

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

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

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

$$\text{F1 - Score (bp)} = \frac{2 \times \text{Accuracy (bp)} \times \text{Completeness (bp)}}{\text{Accuracy (bp)} + \text{Completeness (bp)}}$$

If a bin was composed of contigs derived from different plasmids, then accuracy, completeness and F1-score were reported for each plasmid-bin combination.

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

To evaluate the incorrect inclusion of chromosome-derived contigs into bins, we reported the chromosome contamination metric as specified below.

$$\text{Chromosome contamination} = \frac{\text{Alignment length against chromosome (bp)}}{\text{Total length of predicted bin (bp)}}$$

### Antibiotic Resistance Gene (ARG) Predictions

Resistance genes were predicted by running Abricate (v1.0.1) with the Resfinder [37] database (database indexed on 19 April 2020) using reference plasmids as query, with 80% identity and coverage cut-off. The same software and parameters were used to predict the presence of ARGs in plasmid predictions.

### Evaluating the capacity of the tools to correctly assign ARGs to plasmid predictions

All ARGs assigned to plasmid predictions were classified as plasmid-borne (Detected) or chromosomal (Contamination) by determining the origin of the contig carrying the ARGs, as detailed above. The number of not detected ARGs was obtained by subtracting the number of detected plasmid-borne ARGs in the plasmid predictions from the total number of true plasmid-borne ARGs in reference genomes.

### Exploring plasmidome features

#### Estimation of plasmid copy number

After short-read assembly with unicycler, each contig is assigned a relative coverage value. We used all unitigs that unambiguously aligned to a single replicon to calculate the mean relative coverage of each plasmid. Ambiguous contigs, aligning to more than one location of the genome, were left out of these calculations.

### Identification of repeat elements in genomes

Contigs larger than 1kb that aligned to multiple locations in the genomes were labelled as repeats. These contigs are classified as ‘ambiguous’ by QUAST. Only genomes included in the benchmark dataset were used for this analysis.

### Determining the total length of predicted plasmidome

Total size of predicted plasmidome was obtained by summing up the lengths of all contigs in an isolate predicted as plasmid-derived by plasmidEC.

### Sub-classifying plasmids according to their length

Plasmids were classified as large or small based on size distributions for each species (Supplementary Figure S5). For *E. faecium*, *K. pneumoniae* and *S. enterica* a cut-off value of 18,000 bp was selected, while for *S. aureus* the cut-off was 8,000 bp. For plasmids from less-frequent species, we choose 18,000 bp as a cut-off.

### Querying the SRA database

In order to retrieve the metadata associated with bacterial whole genome sequences uploaded to the SRA database [32], we used the esearch function of the package Entrez Direct (v13.3) [33]. The following search terms were included:

```
esearch -db sra -q ‘(“Bacteria”[Organism] OR “Bacteria Latreille et al. 1825”[Organism]) AND “platform illumina”[Properties] AND (cluster_public[prop] AND “biomol dna”[Properties] AND “strategy wgs”[Properties])’ | efetch -format summary
```

```
esearch -db sra -q ‘(“Bacteria”[Organism] OR “Bacteria Latreille et al. 1825”[Organism]) AND “platform illumina”[Properties] AND (cluster_public[prop] AND “biomol dna”[Properties] AND “strategy wgs”[Properties])’ | efetch -format runinfo
```

The platform term was varied accordingly to include the three sequencing technologies ‘illumina’, ‘Oxford Nanopore’ and ‘PacBio SMRT’.

## Results

### Improving binary classification of contigs for 70 species

We compared the performance of plasmidEC against multiple existing classification tools. For this, we utilized the benchmarking dataset “A” consisting of 405 complete bacterial genomes including 970 plasmids. Of these, 246 genomes and 684 plasmids belonged to species that are highly abundant in public databases (*E. faecium*, *K. pneumoniae*, *S. aureus*, *S. enterica*). The remaining 159 genomes and 286 plasmids belonged to 66 species which are less frequently represented in databases (Supplementary Figure S3, Supplementary data 3).

PlasmidEC achieved the highest F1-Score(contig) values when classifying contigs from *S. enterica* and *E. faecium* (Figure 2, Table 1).

For *S. enterica*, plasmidEC’s F1-Score(contig) is 0.91, with a high recall(contig) (0.95). Notably, the Centrifuge model presented a comparable F1-Score(contig) (0.88) and the highest precision(contig) (0.95). For this species, Platon achieved the most balanced values between recall(contig) (0.86) and precision(contig) (0.87).

For *E. faecium*, the F1-Score(contig) values of plasmidEC (0.96), Platon (0.93) and mlplasmids (0.93) were comparable. However, plasmidEC presented the highest recall(contig) (0.96) of the three, Platon the highest precision(contig) (0.97) and mlplasmids the most balanced metrics, with a precision(contig) of 0.94 and a recall(contig) of 0.93.

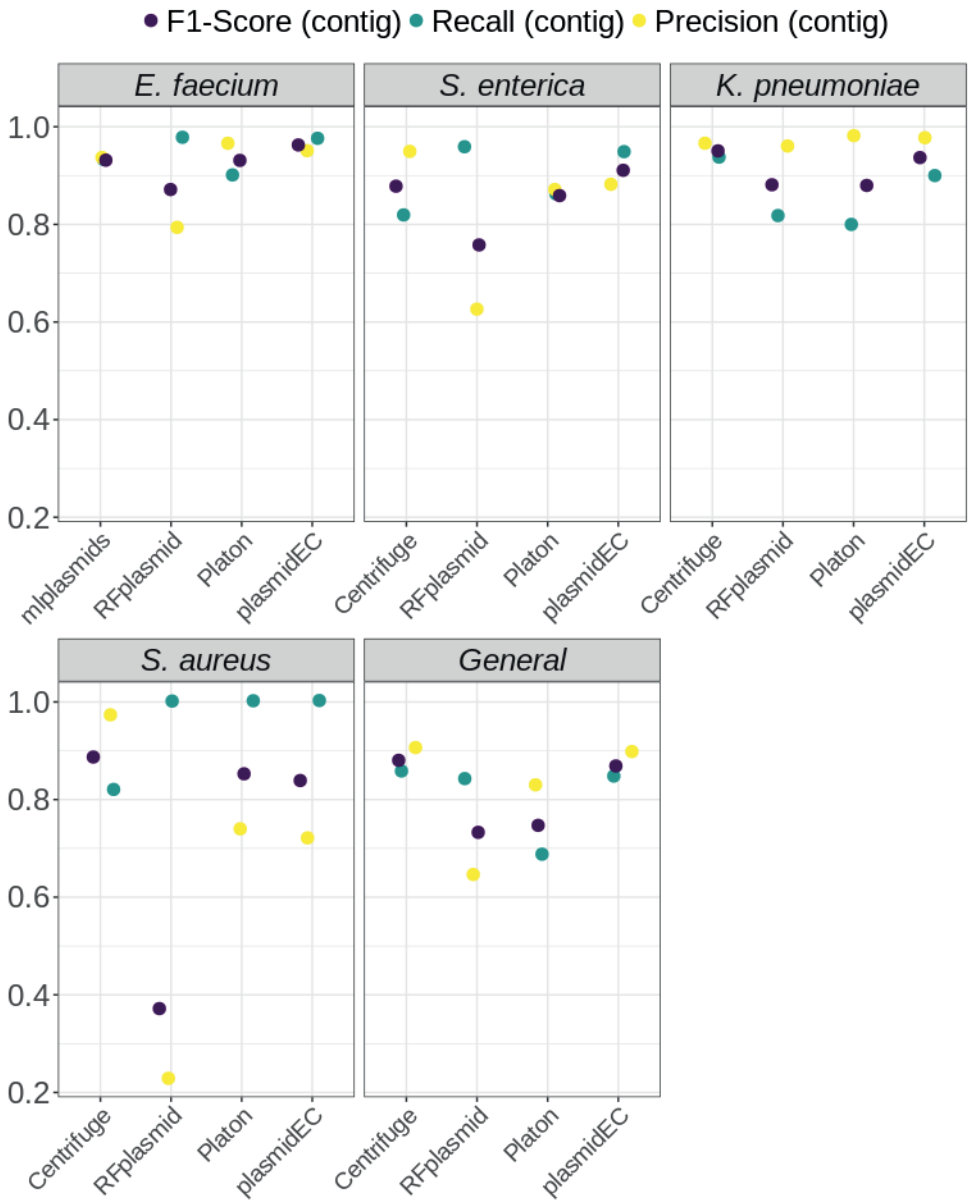
RFPlasmid achieved the highest recall(contig) values in both *S. enterica* en *E. faecium*, but also the lowest precision(contig); this notable imbalance between the metrics resulted in low F1-Score(contig) values for the two species.

The Centrifuge-based classifier had the highest F1-Score(contig) values for *K. pneumoniae*, *S. aureus* and for less-frequent species (indicated as General) (Figure 2, Table 1).

For *K. pneumoniae*, the F1-Score(contig) values obtained by Centrifuge (0.95) and plasmidEC (0.94) were similar, and both tools outperformed RFPlasmid (0.88) and Platon (0.88). Centrifuge presented the highest recall (0.94) and plasmidEC had the highest precision (0.98).

For *S. aureus*, Centrifuge resulted in the highest F1-Score(contig) (0.89) in combination with the highest precision(contig) (0.97), but also in the lowest recall(contig) (0.82). F1-Score(contig) values of PlasmidEC (0.84) and Platon (0.85) were similar, and both tools presented a recall(contig) of 1. Notably, all tools with the exception of Centrifuge, showed precision(contig) values under 0.80 for this species, with RFPlasmid showing the lowest values (0.23).

Finally, for the less frequent species in the dataset, the F1-Score(contig) of Centrifuge (0.88) and plasmidEC (0.87) were comparable, surpassing RFPlasmid (0.73) and Platon (0.75) (Figure 2, Table 1). Centrifuge and PlasmidEC also had the highest recall(contig) of 0.86 and 0.85 respectively, and highest precision(contig) of 0.91 and 0.90 respectively.



**Figure 2.** Performance of different individual binary classification tools (three per species) and of PlasmidEC as an ensemble classifier for binary classification of contigs of multiple species. The ‘General’ category corresponds to species less frequently represented in databases (See Figure S2B for a summary of included species).

**Table 1.** Performance of different individual binary classification tools (three per species) and of PlasmidEC as an ensemble classifier (using input from the three individual tools). The ‘General’ category corresponds to species less frequently represented in databases (See Figure S2B for a summary of the species).

species	software	F1-Score (contig)	Recall (contig)	Precision (contig)
<i>E. faecium</i>	mplasmids	0.93	0.93	0.94
	RFplasmid	0.87	<b>0.98</b>	0.79
	Platon	0.93	0.90	<b>0.97</b>
	plasmidEC	<b>0.96</b>	<b>0.98</b>	0.95
<i>S. enterica</i>	Centrifuge	0.88	0.82	<b>0.95</b>
	RFplasmid	0.76	<b>0.96</b>	0.63
	Platon	0.86	0.86	0.87
	plasmidEC	<b>0.91</b>	0.95	0.88
<i>K. pneumoniae</i>	Centrifuge	<b>0.95</b>	<b>0.94</b>	0.97
	RFplasmid	0.88	0.82	0.96
	Platon	0.88	0.80	<b>0.98</b>
	plasmidEC	0.94	0.90	<b>0.98</b>
<i>S. aureus</i>	Centrifuge	<b>0.89</b>	0.82	<b>0.97</b>
	RFplasmid	0.37	<b>1</b>	0.23
	Platon	0.85	<b>1</b>	0.74
	plasmidEC	0.84	<b>1</b>	0.72
General	Centrifuge	<b>0.88</b>	<b>0.86</b>	<b>0.91</b>
	RFplasmid	0.73	0.84	0.65
	Platon	0.75	0.69	0.83
	plasmidEC	0.87	0.85	0.90

We further compared the performance of plasmidEC and Centrifuge classifiers by evaluating recall(contig) values over individual plasmids. For this, we sub-categorized plasmids into small plasmids (n=410), large non-ARG plasmids (n=311) and large ARG-plasmids (n=215). Notably, across these three groups, the number of contigs into which the plasmids were assembled differed substantially (Supplementary Figure S6). Both tools performed similarly well when identifying contigs of small plasmids, presenting a median recall(contig) of 1 in all species (Supplementary Figure S7). For large ARG-plasmids of *S. enterica*, plasmidEC presented higher recall(contig) values (median=1, IQR=0.92 - 1) than Centrifuge (median=0.83, IQR=0.63 - 1), while for *K. pneumoniae*, *S. aureus* and less-frequent species, the performance of the tools was similar, presenting in all cases a median recall(contig) of 1. Similarly, for large non-ARG plasmids, both tools performed comparably well across most species, with median recall(contig) of 1. However, in the case of *K. pneumoniae*, Centrifuge presented higher recall(contig) values (median=1, IQR=0.96-1) than plasmidEC (median=0.94, IQR=0.83-1).

### Evaluating the performance of gplas

To reconstruct individual plasmids, we selected the best binary classifier in terms of F1-Score(contig) and combined it with gplas. If the F1-Score(contig) was similar between multiple tools (difference < 0.05), then we chose the classifier with the highest recall(contig). This criteria led to combining gplas with Centrifuge for *K. pneumoniae* and for less-frequent species, and with plasmidEC for *S. enterica*, *S. aureus* and *E. faecium* (Supplementary Figure S8).

We compared the performance of gplas against MOB-suite and plasmidSPAdes by applying all tools to the benchmark dataset “B” (Supplementary Figure S3), which consisted of 404 genomes and 953 plasmids. Of these, 158 genomes and 323 plasmids belonged to 75 species less frequently represented in databases, and the remaining 246 genomes and 630 plasmids belonged to *E. faecium*, *K. pneumoniae*, *S. aureus* and *S. enterica*.

Plasmid predictions produced by each tool were aligned to the corresponding complete reference genomes and assessed using the metrics completeness(bp), accuracy(bp) and F1-Score(bp) (Methods). Plasmids were split by size into large and small categories based on cut-offs obtained from the distribution of plasmid sizes (Methods and Supplementary Figure S5). Additionally, since most ARGs are carried by large plasmids (n=2,231, 96.8%) (Supplementary table S1), these were split into large ARG plasmids and large non-ARG plasmids.

When reconstructing large ARG plasmids (n=224), gplas had the highest global F1-Score(bp) value (median=0.76, IQR=0.42 - 0.94), outperforming plasmidSPAdes (median=0.45, IQR=0.17 - 0.90) and MOB-suite (median=0.64, IQR=0.19 - 0.90) (Figure 3A, Table 2). PlasmidSPAdes predictions presented high completeness(bp) values (median=0.83, IQR=0.46 - 0.94), but lacked accuracy(bp) (median=0.46, IQR=0.15 - 0.93) (Figure 3B). In contrast, MOB-suite predictions often showed high accuracy(bp) (median=0.94, IQR= 0.65 - 1), but low completeness(bp) (median= 0.57, IQR= 0.16 - 0.89). Gplas achieved completeness(bp) values (median=0.81, IQR= 0.54 - 0.94) similar to plasmidSPAdes and outperformed MOB-suite in terms of accuracy(bp) (median= 1.00, IQR= 0.53 - 1.00). Similar results were observed when reconstructing large plasmids without resistance genes (n=338) (Supplementary Figure S9 A and B, Table 2).

Table 2. Performance of plasmid reconstruction tools for large ARG- and large non-ARG plasmids.

Type of plasmid	Software	F1-Score (bp) median (IQR)	Completeness (bp) median (IQR)	Accuracy (bp) median (IQR)
Large ARG plasmids	gplas	<b>0.76 (0.42 - 0.94)</b>	0.81 (0.54 - 0.94)	<b>1 (0.53 - 1)</b>
	MOB-suite	0.64 (0.19 - 0.90)	0.57 (0.16 - 0.89)	0.94 (0.65 - 1)
	plasmidSPAdes	0.45 (0.17 - 0.90)	<b>0.83 (0.46 - 0.94)</b>	0.46 (0.15 - 0.93)
Large Non-ARG plasmids	gplas	<b>0.81 (0.39 - 0.99)</b>	0.87 (0.54 - 0.98)	<b>1 (0.52 - 1)</b>
	MOB-suite	0.66 (0.29 - 0.98)	0.63 (0.22 - 0.97)	0.98 (0.72 - 1)
	plasmidSPAdes	0.64 (0.22 - 0.99)	<b>0.95 (0.60 - 1)</b>	0.66 (0.17 - 0.99)

Next, we analysed the reconstruction of large ARG plasmids for each species separately (Figure 3C, Supplementary Table S2). For *S. enterica*, MOB-suite presented the highest F1-Score(bp), whilst for *K. pneumoniae*, *E. faecium* and less-frequent species (Other), gplas predictions resulted in the highest F1-Score(bp) values. Interestingly, for *S. aureus*, all tools performed comparably well, with remarkably high metrics when compared to other species. Notably, gplas presented the most uniform performance across the species, with F1-Score(bp) medians ranging from 0.65 (*K. pneumoniae*) to 1 (*S. aureus*). In contrast, the median F1-Scores(bp) obtained with MOB-suite ranged from 0.48 (*K. pneumoniae*) to 1 (*S. aureus*), while for plasmidSPAdes this metric ranged from 0.2 (*E. faecium*) to 0.99 (*S. aureus*).

Since all metrics are calculated based on plasmids that are detected by each tool (Figure 3C - top), we evaluated the number of detected plasmids per species and per tool. A similar amount of ARG-plasmids were detected by all tools in *E. faecium*, *K. pneumoniae*, *S. aureus* and in less-frequent species (Supplementary table S3). In contrast, plasmidSPAdes only detected 53% (n=21) of *S. enterica* ARG-plasmids, while MOB-suite and gplas detected 97% (n=38) and 94% (n=37), respectively. A total of nine ARG-plasmids could not be detected by any of the tools.

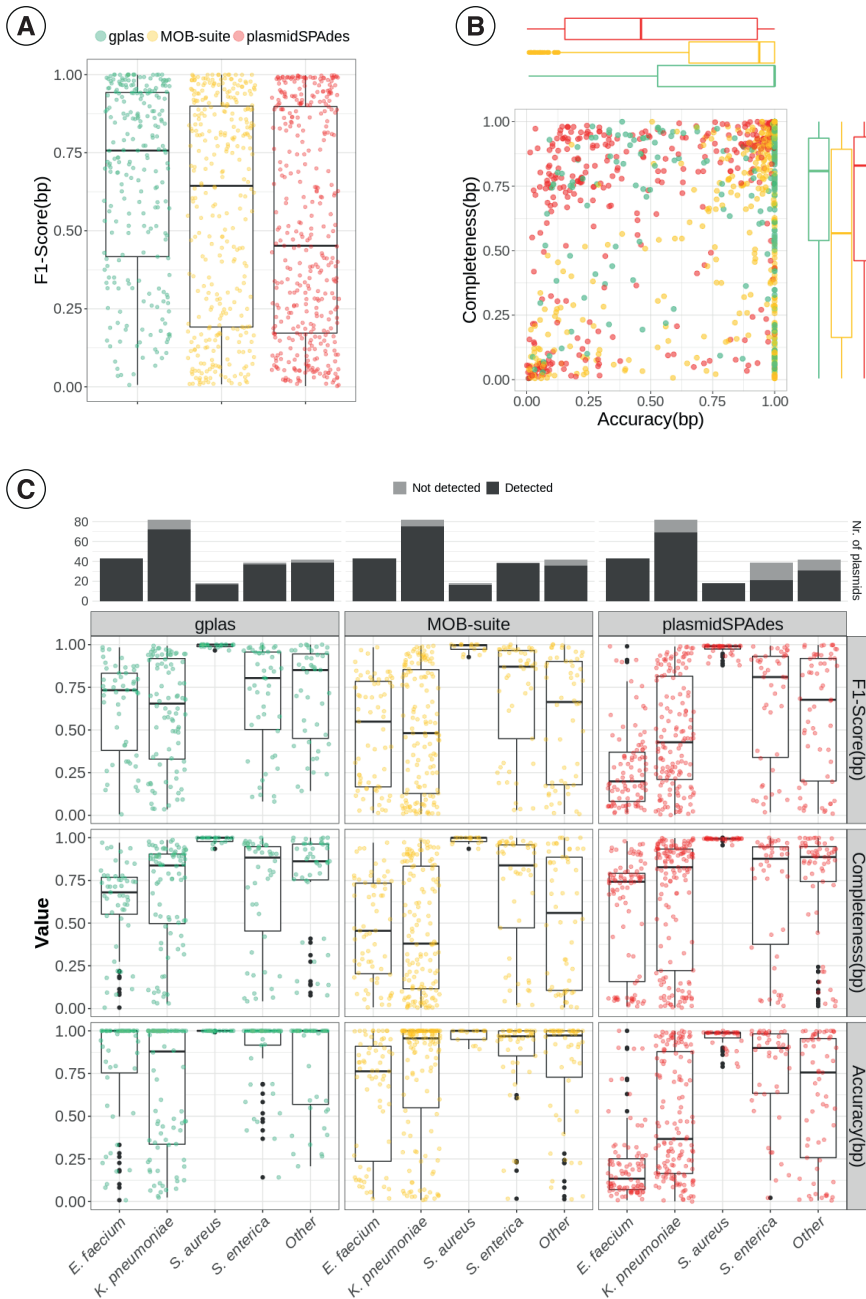
For large plasmids without ARGs, gplas and MOB-suite performed comparably well for all species (Supplementary Table S2, Supplementary Figure S9C - top), while plasmidSPAdes had lower F1-Score(bp) values for *K. pneumoniae* and *E. faecium*. Interestingly, for less-frequent species, MOB-suite detected 59.5% (n=85) of non-ARG plasmids, while gplas and plasmidSPAdes detected 80.3% (n=114). A total of 17 non-ARG plasmids could not be detected by any of the tools.

Next, the ability of each tool to correctly assign plasmid-borne ARGs to predictions was evaluated (Supplementary Figure S10 and Supplementary Table S4). Gplas and MOB-suite performed comparably well in *K. pneumoniae*, *S. enterica* and for less-frequent species, correctly assigning more than 70% of all plasmid-borne ARGs to plasmid predictions. In contrast, plasmidSPAdes correctly assigned lower fractions of plasmid-borne ARGs in these species, ranging from 36.4% (*S. enterica*) to 59.7% (*K. pneumoniae*). In *E. faecium*, all tools correctly assigned at least 80% of all plasmid-borne ARGs, but MOB-suite (90.3%) surpassed gplas (85.0%) and plasmidSPAdes (83.2%). All tools performed best for *S. aureus*, correctly identifying more than 95% of all plasmid-borne ARGs.

We also examined whether plasmid predictions were contaminated with chromosomal sequences. PlasmidSPAdes showed the largest number of predicted plasmid bins contaminated with chromosomal sequences (n=326), and 69.3% (n=226) of these were predominantly composed of chromosomal sequences (chromosome contamination >50%) (Supplementary Figure S11A). Likewise, a total of 167 predictions made by MOB-suite were contaminated and 59.3% (n=155) of these mainly contained chromosomal DNA. Gplas had the lowest number of contaminated bins (n=76) and the majority of these (n=41, 53.9%) showed contamination fractions below 50%. Some differences were observed when analysing each species separately (Supplementary Figure S11B). PlasmidSPAdes had the highest number of contaminated plasmid predictions in *S. aureus* (n=50), *K. pneumoniae* (n=74) and in less-frequent species (n=149). In contrast, MOB-suite had the highest contamination rate in *E. faecium* (n=70). For *S. enterica*, all tools had a similar number of plasmid bins contaminated with chromosomal sequences (range 10 - 20), but gplas included the largest number of plasmid predictions solely composed of chromosomal sequences (n=7).

Finally, we analyzed predictions of small plasmids (n=391). Overall, gplas detected 79.5% of small plasmids, surpassing MOB-suite (70.8%) and PlasmidSpades (71.9%) (Supplementary Figure S12A). Gplas and MOB-suite achieved F1-Score(bp) medians of 1 in most species, except in *E. faecium* in which gplas reached a median of 0.98, while MOB-suite reached a median of 0.74 (Supplementary Figure S12B). PlasmidSPAdes in general had lower F1-Score(bp) values, especially for *K. pneumoniae* and *E. faecium*, which were driven by low completeness(bp) values.





**Figure 3.** Reconstruction metrics from gplas, MOB-suite and plasmidSPAdes. **A**) Overall F1-Score(bp) and **B**) completeness(bp) vs accuracy(bp) for large ARG plasmids from all species. **C**) Number of detected large ARG plasmids per species (top). Reconstruction metrics for detected large ARG plasmids per species (bottom).

## Different species have different plasmidome characteristics which impacts plasmid reconstruction

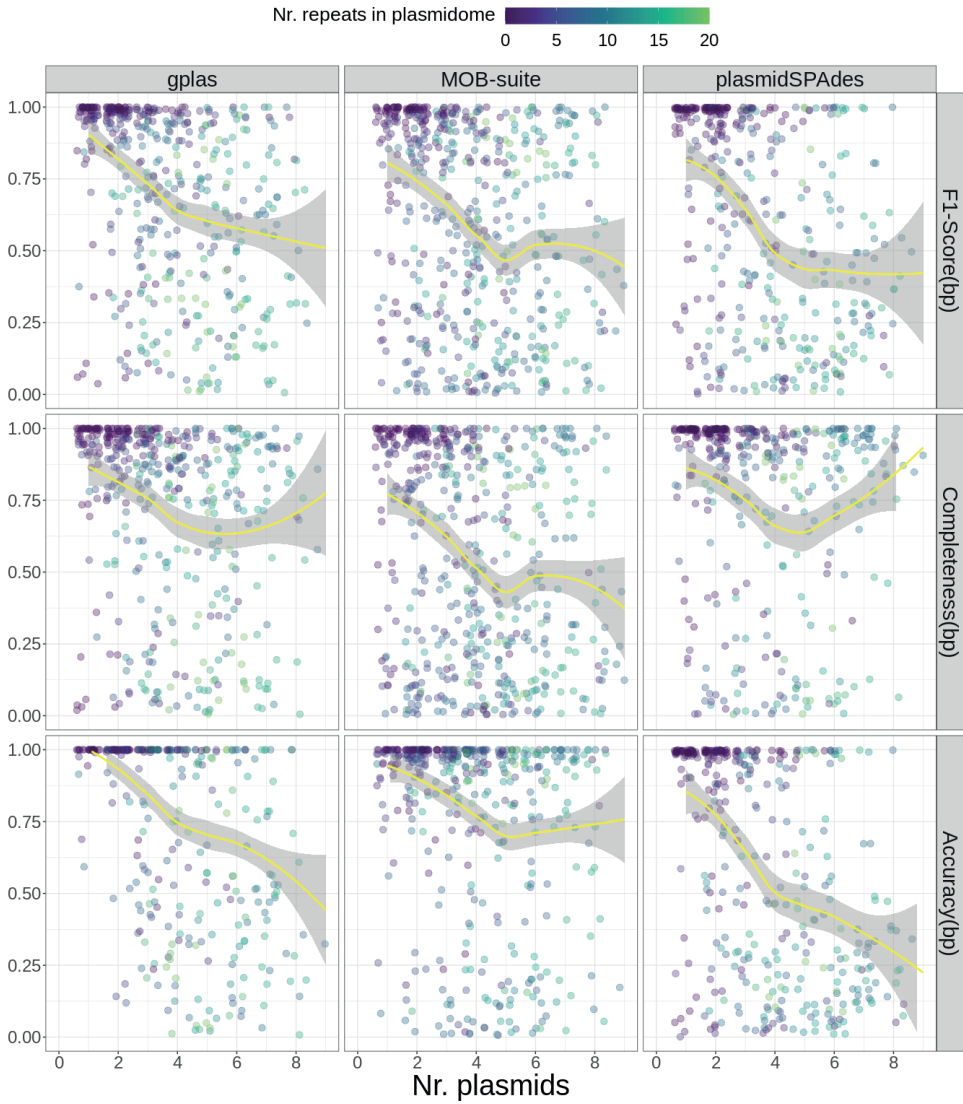
Despite the global performance differences between plasmid reconstruction tools, we observed similar trends in species-level predictive power across plasmid reconstruction tools. All tools had high F1-Score(bp) values (median > 0.8) when reconstructing plasmids from *S. enterica* and *S. aureus*. Conversely, most tools struggled to predict plasmids from *K. pneumoniae* and *E. faecium*. We therefore investigated if differences in the pan-plasmidomes characteristics across species could impact correct plasmid reconstruction. We set to evaluate features as plasmid lengths and copy number, the total number of plasmid per isolate, total plasmidome length and the number of repeated elements.

Using all complete bacterial genomes (n=3,213) uploaded to RefSeq prior to July 26th 2021, (Supplementary data 2), we explored the number of plasmids and the total plasmidome size per genome across species (Supplementary Figure S13). Genomes of *E. faecium* and *K. pneumoniae* had the most plasmids per isolate, with a median of 5 (IQR= 3 - 7) and 3 (IQR=2 - 5) plasmids per genome, respectively. These two species also had the largest plasmidome size, with an average length of 305.3 kb for *K. pneumoniae* and 274.6 kb for *E. faecium*.

*S. aureus* genomes carried a median of 1 (IQR=0-1) plasmid, similarly to *S. enterica* isolates, which also bore a median of 1 plasmid (IQR=0 - 2). Nevertheless, these two species exhibited marked differences in plasmidome lengths, with *S. aureus* plasmidome having an average size of 16.6 kb, while this was 95.0 kb for *S. enterica*.

Using the genomes of *E. faecium*, *K. pneumoniae*, *S. aureus* and *S. enterica* included in the entire benchmark dataset (Supplementary Figure S3), we explored the relation between size and copy number of plasmids (Supplementary Figure S14). We observed an inverse relation between plasmid length and copy number for most species. Similar results were reported for some Enterobacteriaceae genera by Shaw et al. [38]. We also investigated the content of repeat elements in large plasmids (Supplementary Figure S15). *E. faecium* plasmids contained the most repeats, with a median of 4 (IQR= 2 - 10), followed by *K. pneumoniae* plasmids that carried a median of 3 (IQR= 0 - 5) repeats, with some plasmids having up to 22 repeats. *S. enterica* plasmids carried a median of 1 repeat (IQR=0-3). Notably, plasmids of *S. aureus* did not contain any repeat elements.

Finally, using genomes included in benchmark dataset “B” (Supplementary Figure S3), we explored the relation between reconstruction metrics and the plasmidome features previously described. We plotted F1-Score(bp), completeness(bp) and accuracy(bp) values for the predictions of large plasmids as a function of the number of plasmids in an isolate and fitted a LOESS regression (alpha=0.8) (Figure 4). All tools appeared to perform worse on all metrics, with increasing number of plasmids. When compared to other tools, MOB-suite predictions decayed more rapidly in completeness(bp) when the number of plasmids increased. In contrast, plasmidSPAdes and gplas predictions showed a stronger decay in accuracy(bp). A similar trend in the decay of F1-Score(bp) was observed when either the number of repeats in a replicon or the total plasmidome size were plotted against the reconstruction metrics (Supplementary Figure S16 and Supplementary Figure S17). Interestingly, plasmid copy number significantly impacted the quality of plasmidSPAdes predictions (Supplementary Figure S18). In particular, plasmids that had a copy number close to one were poorly reconstructed, displaying very low values of completeness(bp). In contrast, MOB-suite and gplas predictions were less affected by plasmid copy number.



**Figure 4.** Reconstruction metrics for large plasmids (n=562) plotted as a function of the number of plasmids per isolate. The data is colored according to the total number of repeats present in the plasmidome of the isolate. Yellow line indicates a LOESS regression ( $\alpha=0.8$ ).

## Discussion

In this work, we developed two novel methods to predict plasmid contigs of any bacterial species. We compared the performance of these methods (Centrifuge and plasmidEC) against other binary classification tools. Later, we used gplas (combined with plasmidEC or Centrifuge) to reconstruct plasmids of more than 70 bacterial species and benchmarked it against MOB-suite and plasmidSPAdes by applying the tools to a dataset consisting of 404 complete bacterial genomes and 953 plasmids. We found that gplas performed consistently well when reconstructing large ARG-plasmids in multiple species, in contrast to a varying performance of the other tools. For non-ARG-plasmids, the differences in performance were less pronounced.

PlasmidEC and Centrifuge outperformed other existing binary classifiers for all evaluated species. Since Centrifuge relies on a database to classify contigs, the good performance of this tool appears to suggest that most of the bacterial pan-plasmidome diversity is appropriately captured by the plasmid sequences currently available in databases. However, this result could also be a consequence of solely selecting sequences from public databases, such as RefSeq and SRA, as a benchmark dataset. These databases are known to preferentially contain bacterial sequences from clinical environments such as species outlined on WHO priority pathogens list, and from a limited number of geographical origins [39]. The inclusion of novel sequences, from less-frequent sources or other geographical locations, in a future benchmark study, could shed light onto the generalizability of the performance of the methods developed here.

An important limitation of most classification tools is their binary output, namely contigs are forced to be classified as either plasmid- or chromosome-derived. This feature complicates the prediction of mobile genetic elements, which can be carried by both plasmids and chromosomes [21,40,41]. To partially overcome this limitation, the centrifuge-based classifiers described in this paper can assign contigs to an ‘unclassified’ category if they align to the plasmidome and chromosome fraction of the database in similar proportions. Despite this, classifying ambiguous contigs without exploration of their flanking sequences in each particular genome could frequently lead to misclassification. Recently, the authors of plASgraph [42] and 3CAC [43] demonstrated that exploring the assembly graph through convolutional neural networks improves the identification of plasmid contigs in both WGS and metagenomics data. Consequently, integrating plASgraph into plasmidEC could improve the classification of these ambiguous contigs leading to higher recall and precision values.

When reconstructing individual ARG-plasmids of *S. enterica* and *S. aureus*, gplas, MOB-suite and plasmidSPAdes performed comparably well, displaying similar values of completeness(bp) and accuracy(bp). However, gplas considerably outperformed the other tools when reconstructing ARG-plasmids of *E. faecium* and *K. pneumoniae*. For these species, predictions generated by MOB-suite had lower completeness(bp) values while plasmidSPAdes predictions lacked accuracy(bp). Notably, *E. faecium* and *K. pneumoniae* pan-plasmidomes seemed more complex when compared to those of other species, presenting a larger number of plasmids per isolate and more repeats per plasmid. These genomic characteristics are expected to cause more fragmented and entangled assemblies, which would lead to difficulties when attempting to reconstruct plasmids with tools that uniquely rely on either reference- or graph-based approaches, as MOB-suite and plasmidSPAdes do. The observation that gplas predictions were less affected by these genomic fea-

tures further demonstrates the benefits of combining assembly graph information with accurate contig classifications to predict individual plasmids. It must be noted that the MOB-suite database was last updated in May 2020, when the total number of publicly available complete genomes was lower. While it would be interesting to evaluate the performance of MOB-suite coupled to an updated database, we previously demonstrated for *E. coli* (Chapter 3) that the main limitation of MOB-suite was not its capacity to detect plasmid-derived contigs, but its ability to correctly bin multiple contigs together into individual predictions. Consequently, it is unlikely that a newer database would lead to substantial improvements in the performance of MOB-suite.

When reconstructing large ARG-plasmids from species with low abundance in databases, gplas had the highest F1-Score(bp) (median=0.85, IQR=0.45-0.95), outperforming MOB-suite and plasmidSPAdes. This means that gplas, combined with our species-independent Centrifuge model, could be used to explore the whole bacterial pan-plasmidome diversity currently available in the SRA database. Additionally, these results indicate that both tools could potentially be applied to metagenomic samples. However, the application of gplas to metagenomes would require implementing a number of computational solutions aimed at reducing memory and time requirements, such as parallelization of the plasmid walks and limiting the maximum number of nodes explored in each walk. Moreover, because metagenomes contain sequences from multiple genomes with different abundances, the variation in sequencing coverage of chromosome-derived nodes will not be a useful metric for constructing plasmid-walks.

Even with the successful reconstruction of plasmids from multiple species by gplas, some limitations remain. First, plasmids that are represented as disconnected nodes in the assembly graph (degree 0) are currently assigned to the ‘unbinned’ category. These isolated nodes could represent linear plasmids, which occur in multiple species [44–49], but also plasmids that were not assembled correctly or were not fully sequenced (Chapter 3). The ‘unbinned’ category also includes plasmid-predicted nodes that are surrounded by only chromosomal nodes or by nodes that exhibit large sequencing coverage variations. A simple solution will be to subclassify unbinned nodes into different categories, namely disconnected components and potential chromosomal contamination. This will allow the user to select which predicted plasmid bins are to be included or excluded from the analysis. If many unbinned nodes are observed, different assembly tools and parameters should be tested to improve plasmid reconstructions. Second, although the performance of our method was similar across species, reconstructing individual plasmids in genomes with complex plasmidomes would still pose a challenge, especially when an isolate holds multiple plasmids with similar copy numbers that share repeated elements. In these cases, having an *a priori* estimation of the number of distinct large plasmids carried by an isolate could also help in fine-tuning the parameters used by gplas to partition the plasmidome network more accurately. Plasmid number could be estimated using a combination of the number of incompatibility groups, relaxases, origins of replication present in the isolate.

In conclusion, in this work we showed that gplas, combined with a robust binary classification tool, constitutes the best available method to reconstruct plasmids from a wide range of bacterial species in the absence of long-read data.

**References**

1. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. 2022;399: 629–655.
2. Terreni M, Taccani M, Pregnolato M. New Antibiotics for Multidrug-Resistant Bacterial Strains: Latest Research Developments and Future Perspectives. *Molecules*. 2021;26: 2671.
3. 2020 Antibacterial agents in clinical and preclinical development: an overview and analysis. World Health Organization; 2021.
4. Monoclonal antibodies as antibacterial therapies: thinking outside of the box. *Lancet Infect Dis*. 2021;21: 1201–1202.
5. Motley MP, Banerjee K, Fries BC. Monoclonal antibody-based therapies for bacterial infections. *Curr Opin Infect Dis*. 2019;32: 210–216.
6. Uytendaele S, Chen B, Onsea J, Ruythooren F, Debaveye Y, Devolder D, et al. Safety and efficacy of phage therapy in difficult-to-treat infections: a systematic review. *Lancet Infect Dis*. 2022;22: e208–e220.
7. Van Nieuwenhuysse B, Van der Linden D, Chatzis O, Lood C, Wagemans J, Lavigne R, et al. Bacteriophage-antibiotic combination therapy against extensively drug-resistant *Pseudomonas aeruginosa* infection to allow liver transplantation in a toddler. *Nature Communications*. 2022. doi:10.1038/s41467-022-33294-w
8. Bokhary H, Pangesti KNA, Rashid H, El Ghany MA, Hill-Cawthorne GA. Travel-Related Antimicrobial Resistance: A Systematic Review. *Tropical Medicine and Infectious Disease*. 2021;6. doi:10.3390/tropical-med6010011
9. Irfan M, Almotiri A, AlZeyadi ZA. Antimicrobial Resistance and Its Drivers-A Review. *Antibiotics* (Basel, Switzerland). 2022;11. doi:10.3390/antibiotics11101362
10. Environmental antimicrobial resistance and its drivers: a potential threat to public health. *Journal of Global Antimicrobial Resistance*. 2021;27: 101–111.
11. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat Rev Microbiol*. 2021;19: 347–359.
12. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun*. 2020;11: 1–13.
13. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *eLife*. 2020. doi:10.7554/elife.53886
14. Yamagishi T, Matsui M, Sekizuka T, Ito H, Fukusumi M, Uehira T, et al. A prolonged multispecies outbreak of IMP-6 carbapenemase-producing Enterobacterales due to horizontal transmission of the IncN plasmid. *Sci Rep*. 2020;10: 4139.
15. de Man TJB, Yaffee AQ, Zhu W, Batra D, Alyanak E, Rowe LA, et al. Multispecies Outbreak of Verona Integron-Encoded Metallo- $\beta$ -Lactamase-Producing Multidrug-Resistant Bacteria Driven by a Promiscuous Incompatibility Group A/C2 Plasmid. *Clin Infect Dis*. 2020;72: 414–420.
16. Hidalgo L, de Been M, Rogers MRC, Schürch AC, Scharringa J, van der Zee A, et al. Sequence-based epidemiology of an OXA-48 plasmid during a hospital outbreak. *Antimicrob Agents Chemother*. 2019. doi:10.1128/AAC.01204-19
17. Bosch T, Lutgens SPM, Hermans MHA, Wever PC, Schneeberger PM, Renders NHM, 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.
18. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics*. 2017;3. doi:10.1099/mgen.0.000132

19. Arredondo-Alonso S, Pöntinen AK, Cléon F, Gladstone RA, Schürch AC, Johnsen PJ, et al. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience*. 2021;10. doi:10.1093/gigascience/giab079
20. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial genomics*. 2017;3. doi:10.1099/mgen.0.000128
21. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom*. 2020;6. doi:10.1099/mgen.0.000398
22. Royer G, Decusser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000211
23. van der Graaf-van Bloois L, Wagenaar JA, Zomer AL. RFPPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microb Genom*. 2021;7. doi:10.1099/mgen.0.000683
24. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics*. 2020;36: 3874–3876.
25. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000206
26. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis*. 2019;19: 56–66.
27. Lim C, Takahashi E, Hongsuwan M, Wuthiekanun V, Thamlikitkul V, Hinjoy S, et al. Epidemiology and burden of multidrug-resistant bacterial infection in a developing country. *eLife*. 2016. doi:10.7554/elife.18082
28. Pöntinen AK, Top J, Arredondo-Alonso S, Tonkin-Hill G, Freitas AR, Novais C, et al. Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. *Nat Commun*. 2021;12: 1–13.
29. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000224
30. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 2016;26: 1721–1729.
31. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016;32: 3380–3387.
32. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res*. 2011;39: D19.
33. Kans J. Entrez Direct: E-utilities on the Unix Command Line. Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US); 2023.
34. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17: 1–14.
35. 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.
36. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29: 1072–1075.
37. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother*. 2020;75: 3491–3500.
38. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pan-genome of wastewater- and livestock-associated Enterobacteriaceae. *Sci Adv*. 2021;7. doi:10.1126/sciadv.abe3868

39. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol.* 2021;19: e3001421.
40. Wang Y, Batra A, Schulenburg H, Dagan T. Gene sharing among plasmids and chromosomes reveals barriers for antibiotic resistance gene transfer. *Philos Trans R Soc Lond B Biol Sci.* 2022;377: 20200467.
41. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, 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. doi:10.1073/pnas.2008731118
42. Sielemann J, Sielemann K, Brejová B, Vinař T, Chauve C. plASgraph - using graph neural networks to detect plasmid contigs from an assembly graph. *bioRxiv.* 2022. p. 2022.05.24.493339. doi:10.1101/2022.05.24.493339
43. Pu L, Shamir R. 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics.* 2022;38: ii56–ii61.
44. Kinashi H. Giant linear plasmids in *Streptomyces*: a treasure trove of antibiotic biosynthetic clusters. *J Antibiot.* 2010;64: 19–25.
45. Plasterk RH, Simon MI, Barbour AG. Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrelia hermsii*. *Nature.* 1985;318. doi:10.1038/318257a0
46. Hayakawa T, Tanaka T, Sakaguchi K, Otake N, Yonehara H. A linear plasmid-like DNA in *Streptomyces sp.* Producing lankacidin group antibiotics. *J Gen Appl Microbiol.* 1979;25: 255–260.
47. Meinhardt F, Schaffrath R, Larsen M. Microbial linear plasmids. *Appl Microbiol Biotechnol.* 1997;47. doi:10.1007/s002530050936
48. Hawkey J, Cottingham H, Tokolyi A, Wick RR, Judd LM, Cerdeira L, et al. Linear plasmids in *Klebsiella* and other Enterobacteriaceae. *Microbial Genomics.* 2022;8: 000807.
49. Boumasmoud M, Dengler HV, Schweizer TA, Meyer L, Chakrakodi B, Schreiber PW, et al. Genomic Surveillance of Vancomycin-Resistant *Enterococcus faecium* Reveals Spread of a Linear Plasmid Conferring a Nutrient Utilization Advantage. *MBio.* 2022;13. doi:10.1128/mbio.03771-21



## Supplementary Materials

### Supplementary Data

Supplementary data can be downloaded from: <https://doi.org/10.5281/zenodo.7926600>

### Supplementary Tables

**Supplementary Table S1.** Number of antibiotic resistance genes in plasmids from the entire benchmark dataset, categorized by plasmid size and species.

species	Plasmid size classification	Nr. of plasmids	Nr. of ARGs
<i>E. faecium</i>	large	156	241
	small	145	1
<i>K. pneumoniae</i>	large	407	1076
	small	328	22
Other	large	346	363
	small	263	20
<i>S. aureus</i>	large	49	53
	small	22	5
<i>S. enterica</i>	large	144	498
	small	63	25

## Chapter 4

**Supplementary Table S2.** Performance metrics of plasmid reconstruction tools for large ARG- and non-ARG plasmids for different species. The 'Other' category includes 75 species less-frequently represented in databases.

Plasmid Type	Software	Species	Completeness median (IQR)	Accuracy median (IQR)	F1-Score median (IQR)
Large ARG-plasmid	gplas	<i>E. faecium</i>	0.68 (0.55 - 0.77)	1 (0.75 - 1)	0.73 (0.38 - 0.83)
		<i>K. pneumoniae</i>	0.84 (0.5 - 0.9)	0.88 (0.34 - 1)	0.65 (0.33 - 0.92)
		Other	0.86 (0.75 - 0.96)	1 (0.57 - 1)	0.85 (0.45 - 0.95)
		<i>S. aureus</i>	1 (0.98 - 1)	1 (1 - 1)	1 (0.99 - 1)
		<i>S. enterica</i>	0.88 (0.45 - 0.95)	1 (0.92 - 1)	0.8 (0.5 - 0.96)
	MOB-suite	<i>E. faecium</i>	0.46 (0.2 - 0.73)	0.76 (0.24 - 0.91)	0.55 (0.17 - 0.78)
		<i>K. pneumoniae</i>	0.38 (0.12 - 0.83)	0.96 (0.55 - 1)	0.48 (0.13 - 0.85)
		Other	0.56 (0.11 - 0.89)	0.97 (0.73 - 1)	0.66 (0.18 - 0.9)
		<i>S. aureus</i>	1 (0.98 - 1)	1 (0.95 - 1)	1 (0.97 - 1)
		<i>S. enterica</i>	0.84 (0.47 - 0.96)	0.97 (0.85 - 1)	0.87 (0.45 - 0.97)
	plasmidSPAdes	<i>E. faecium</i>	0.74 (0.16 - 0.79)	0.13 (0.07 - 0.25)	0.2 (0.08 - 0.37)
		<i>K. pneumoniae</i>	0.83 (0.22 - 0.93)	0.37 (0.16 - 0.88)	0.43 (0.21 - 0.82)
		Other	0.89 (0.74 - 0.95)	0.76 (0.26 - 0.96)	0.68 (0.2 - 0.92)
		<i>S. aureus</i>	0.99 (0.99 - 1)	0.99 (0.96 - 0.99)	0.99 (0.97 - 0.99)
		<i>S. enterica</i>	0.88 (0.38 - 0.95)	0.9 (0.63 - 0.98)	0.81 (0.34 - 0.93)
Large nonARG-plasmid	gplas	<i>E. faecium</i>	0.64 (0.39 - 0.74)	0.73 (0.28 - 0.95)	0.53 (0.27 - 0.73)
		<i>K. pneumoniae</i>	0.86 (0.59 - 0.97)	0.97 (0.48 - 1)	0.76 (0.38 - 0.97)
		Other	0.94 (0.39 - 0.99)	1 (0.57 - 1)	0.96 (0.42 - 1)
		<i>S. aureus</i>	0.99 (0.96 - 1)	1 (1 - 1)	1 (0.98 - 1)
		<i>S. enterica</i>	0.99 (0.95 - 1)	1 (1 - 1)	0.99 (0.97 - 1)
	MOB-suite	<i>E. faecium</i>	0.43 (0.16 - 0.68)	0.68 (0.26 - 0.84)	0.44 (0.2 - 0.66)
		<i>K. pneumoniae</i>	0.44 (0.17 - 0.93)	0.97 (0.72 - 1)	0.56 (0.25 - 0.94)
		Other	0.92 (0.34 - 1)	1 (0.91 - 1)	0.9 (0.48 - 1)
		<i>S. aureus</i>	0.95 (0.51 - 0.99)	1 (1 - 1)	0.98 (0.65 - 1)
		<i>S. enterica</i>	0.99 (0.94 - 1)	1 (0.98 - 1)	0.98 (0.95 - 1)
	plasmidSPAdes	<i>E. faecium</i>	0.45 (0.18 - 0.85)	0.16 (0.07 - 0.31)	0.22 (0.09 - 0.38)
		<i>K. pneumoniae</i>	0.89 (0.59 - 0.98)	0.43 (0.19 - 0.96)	0.52 (0.26 - 0.92)
		Other	0.98 (0.74 - 1)	0.98 (0.22 - 1)	0.93 (0.22 - 1)
		<i>S. aureus</i>	0.97 (0.92 - 1)	0.99 (0.98 - 0.99)	0.98 (0.95 - 0.99)
		<i>S. enterica</i>	0.99 (0.91 - 1)	0.99 (0.86 - 1)	0.97 (0.56 - 1)

## Accurately reconstructing AMR plasmids of multiple bacterial species from short reads

**Supplementary Table S3.** Number of detected (and not-detected) large plasmids per tool and species. A plasmid was considered detected if a single (non-ambiguous) contig is included into the plasmid predictions. The 'Other' category includes 75 species less-frequently represented in databases.

Plasmid Type	Species	variable	Detected	Not detected	
Large ARG-plasmid	<i>E. faecium</i>	gplas	43	0	
		MOB-suite	43	0	
		plasmidSPAdes	43	0	
	<i>K. pneumoniae</i>	gplas	72	10	
		MOB-suite	75	7	
		plasmidSPAdes	69	13	
	<i>S. aureus</i>	gplas	17	1	
		MOB-suite	16	2	
		plasmidSPAdes	18	0	
	<i>S. enterica</i>	gplas	37	2	
		MOB-suite	38	1	
		plasmidSPAdes	21	18	
	Other	gplas	39	3	
		MOB-suite	36	6	
		plasmidSPAdes	31	11	
	Large nonARG-plasmid	<i>E. faecium</i>	gplas	34	5
			MOB-suite	35	4
			plasmidSPAdes	38	1
<i>K. pneumoniae</i>		gplas	104	15	
		MOB-suite	115	4	
		plasmidSPAdes	103	16	
<i>S. aureus</i>		gplas	6	2	
		MOB-suite	6	2	
		plasmidSPAdes	8	0	
<i>S. enterica</i>		gplas	23	7	
		MOB-suite	22	8	
		plasmidSPAdes	19	11	
Other		gplas	114	28	
		MOB-suite	85	57	
		plasmidSPAdes	114	28	

## Chapter 4

**Supplementary Table S4.** Number of detected (and not-detected) antibiotic resistance genes in plasmid predictions generated by gplas, MOB-suite and plasmidSPAdes, using benchmark dataset “B”. The number of ARGs of chromosomal origin, incorrectly included in predictions, are also detailed. The ‘Other’ category includes 75 species less-frequently represented in databases.

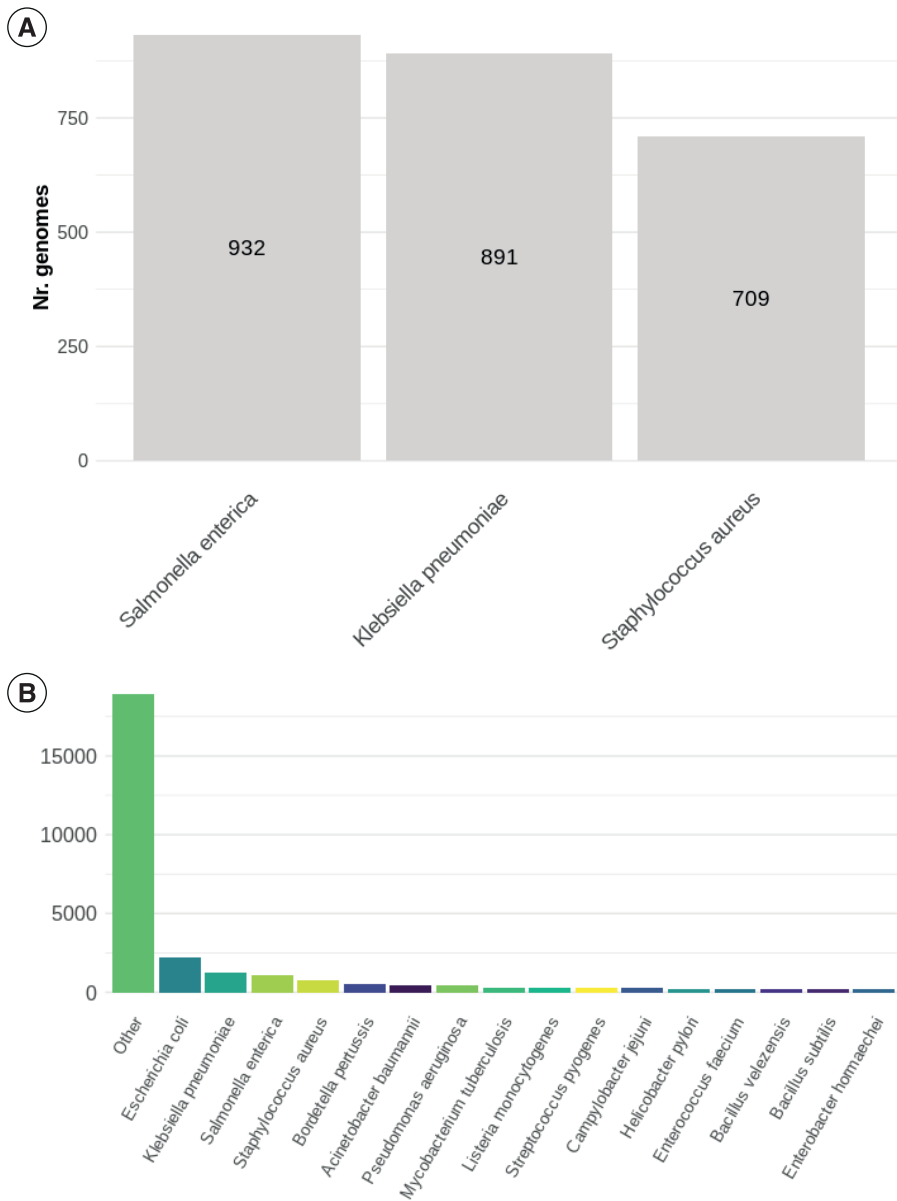
Species	Replicon origin	Classification	Software	Nr. ARGs	Percentage
<i>E. faecium</i>	Plasmid-borne	Detected	gplas	96	84.96
			MOB-suite	102	90.27
			plasmidSPAdes	94	83.19
	Not detected	gplas	17	15.04	
		MOB-suite	11	9.73	
		plasmidSPAdes	19	16.81	
	Chromosomal	Contamination	gplas	16	Not applicable
			MOB-suite	24	Not applicable
			plasmidSPAdes	8	Not applicable
<i>K. pneumoniae</i>	Plasmid-borne	Detected	gplas	372	81.05
			MOB-suite	398	86.71
			plasmidSPAdes	274	59.69
	Not detected	gplas	87	18.95	
		MOB-suite	61	13.29	
		plasmidSPAdes	185	40.31	
	Chromosomal	Contamination	gplas	11	Not applicable
			MOB-suite	32	Not applicable
			plasmidSPAdes	8	Not applicable
<i>S. enterica</i>	Plasmid-borne	Detected	gplas	197	74.62
			MOB-suite	225	85.23
			plasmidSPAdes	96	36.36
	Not detected	gplas	67	25.38	
		MOB-suite	39	14.77	
		plasmidSPAdes	168	63.64	
	Chromosomal	Contamination	gplas	47	Not applicable
			MOB-suite	58	Not applicable
			plasmidSPAdes	8	Not applicable
<i>S. aureus</i>	Plasmid-borne	Detected	gplas	32	100
			MOB-suite	31	96.88
			plasmidSPAdes	32	100
	Not detected	gplas	0	0	
		MOB-suite	1	3.12	
		plasmidSPAdes	0	0	
	Chromosomal	Contamination	gplas	3	Not applicable
			MOB-suite	6	Not applicable
			plasmidSPAdes	1	Not applicable

Accurately reconstructing AMR plasmids of multiple bacterial species from short reads

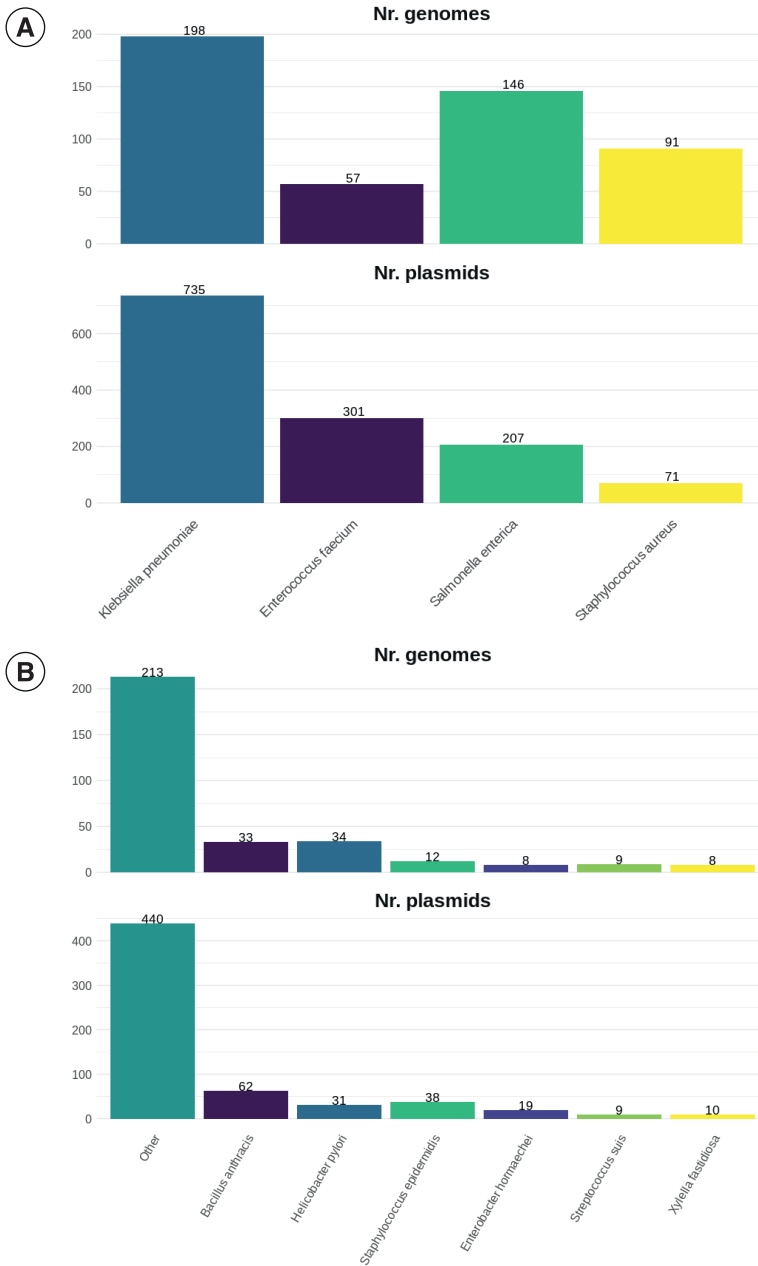
Supplementary Table S4. Cont.

Species	Replicon origin	Classification	Software	Nr. ARGs	Percentage
<i>Other</i>	Plasmid-borne	Detected	gplas	149	74.5
			MOB-suite	143	71.5
			plasmidSPAdes	110	55
	Not detected		gplas	51	25.5
			MOB-suite	57	28.5
			plasmidSPAdes	90	45
	Chromosomal	Contamination	gplas	27	Not applicable
			MOB-suite	22	Not applicable
			plasmidSPAdes	51	Not applicable

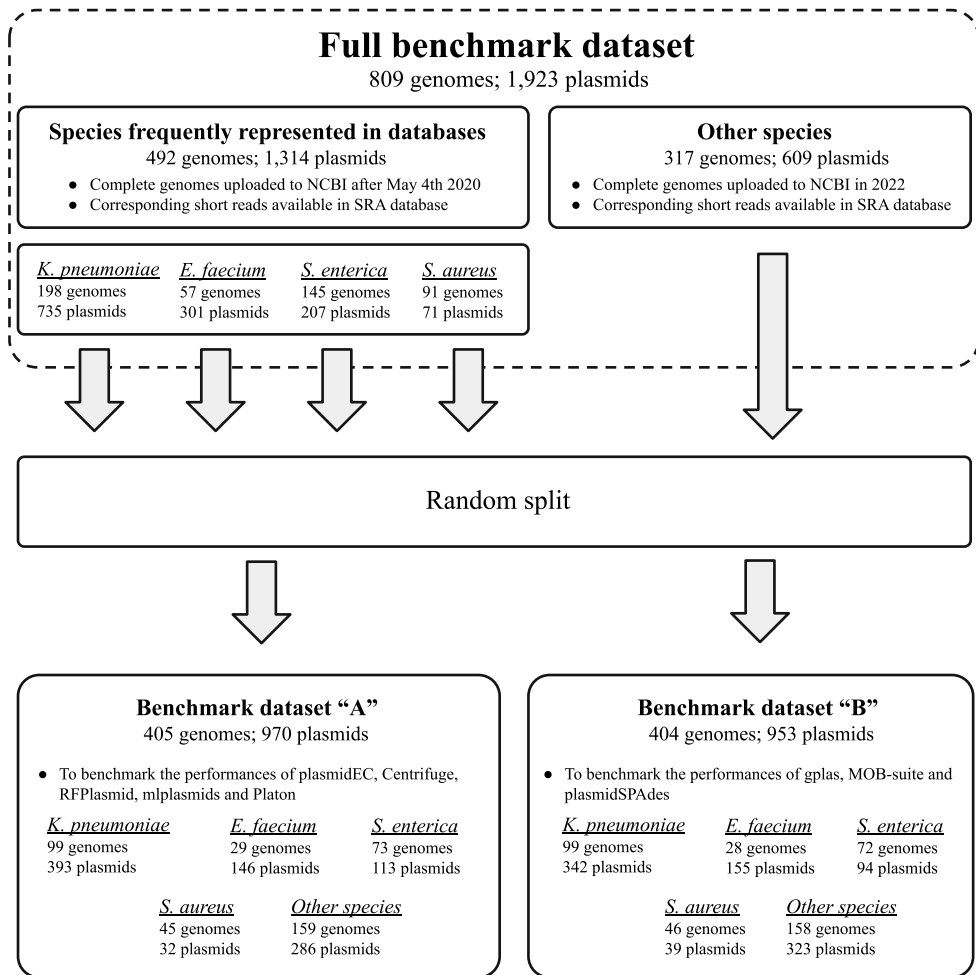
## Supplementary Figures



**Supplementary Figure S1.** A) Number of genomes included in the construction of species-specific Centrifuge databases. B) Genomes included in the construction of the species-independent Centrifuge database. The 'Other' category includes genomes from species of which less than 200 genomes were present in RefSeq.

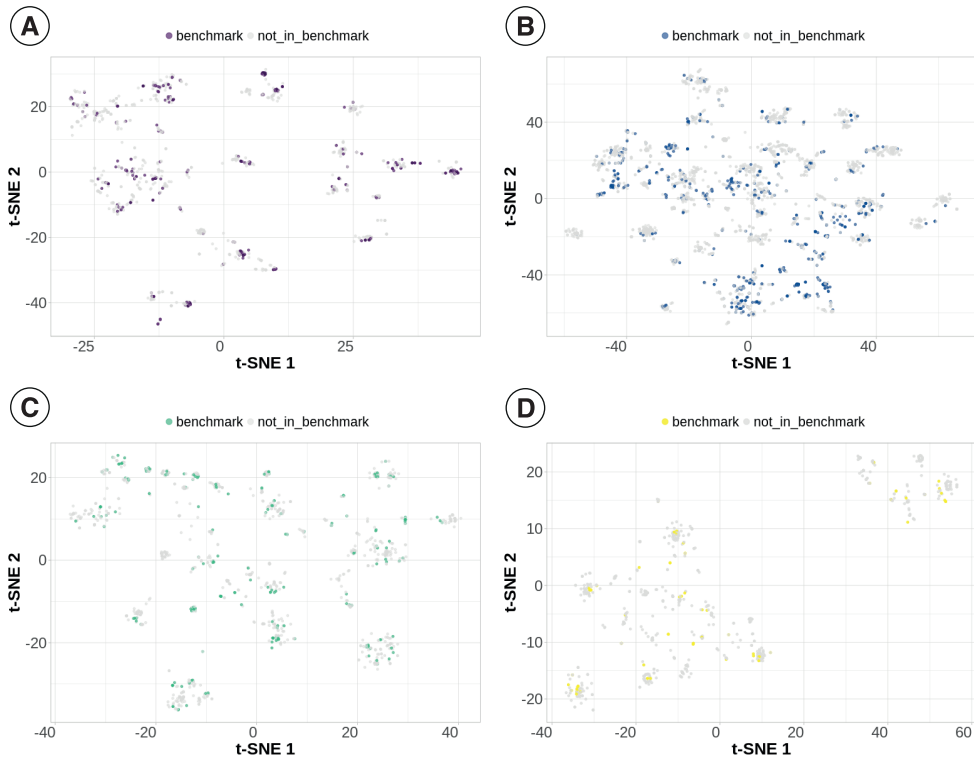


**Supplementary Figure S2.** A) Number of genomes and plasmids from common human pathogens, highly abundant in databases, included in the benchmark dataset. B) Number of genomes and plasmids of less-frequent species included in the benchmark dataset. The ‘Other’ category includes species with less than 10 genomes in the dataset. The complete dataset was fractioned in two equally-sized groups to independently test performances of plasmidEC and gplas. A more detailed description of which genome was included in each group can be found in Supplementary Data 2 and in Supplementary Figure S3.

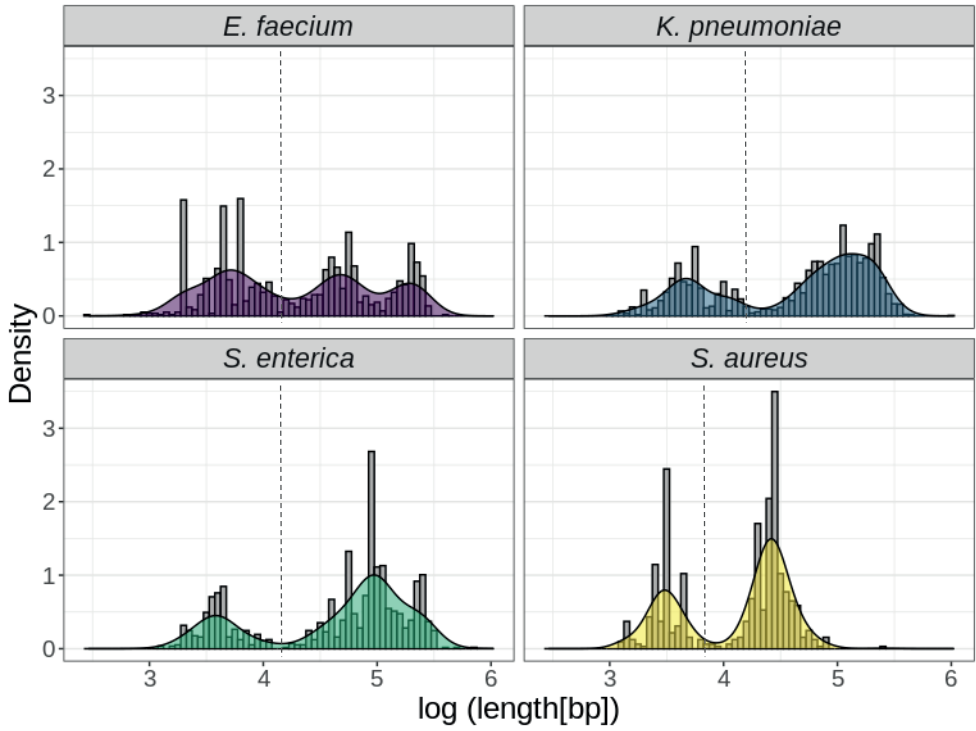


**Supplementary Figure S3.** Composition of benchmark datasets. Dataset “A” was used to benchmark the performance of binary classifiers, while dataset “B” was used to benchmark plasmid reconstruction tools performances.

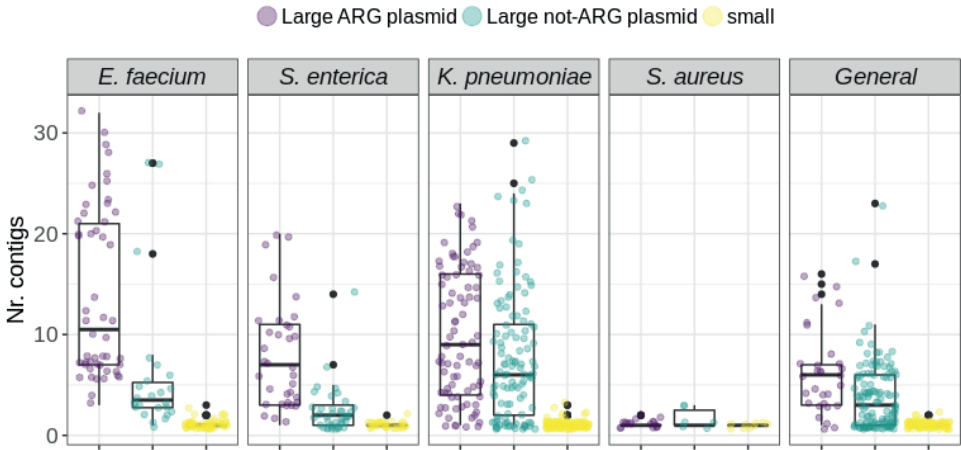




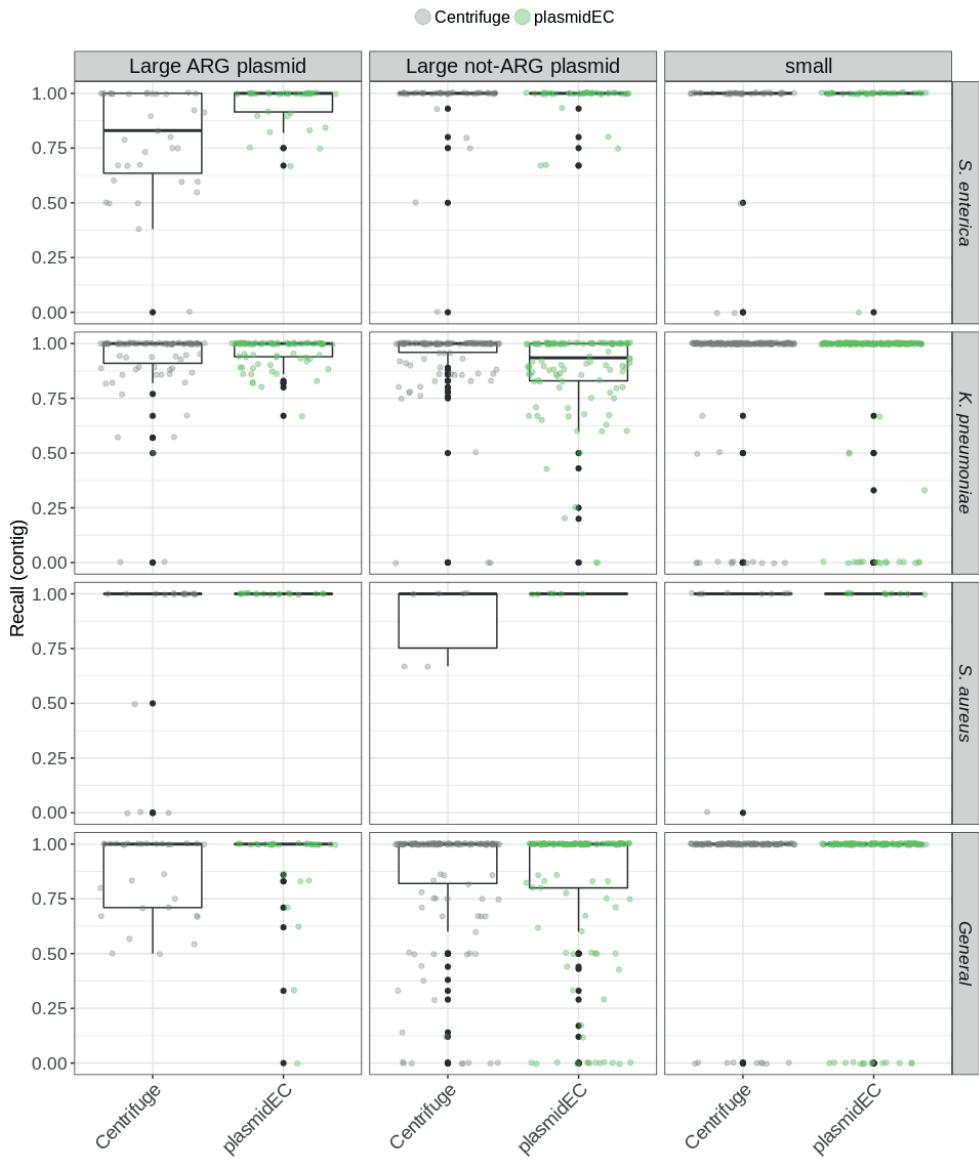
**Supplementary Figure S4.** t-SNE visualization of MASH distances ( $s=10,000$ ,  $k=21$ ) between all complete plasmid sequences in RefSeq (accession date: July 26th, 2021) for (A) *E. faecium*, (B) *K. pneumoniae*, (C), *S. enterica* and (D) *S. aureus*. Plasmids that are colored are included in the benchmark datasets for the respective species-specific models.



**Supplementary Figure S5.** Plasmid size distribution per species. Dotted lines represent cut-offs selected for classifying plasmids according to size into either 'small' or 'large'. The y-axis shows Kernell probability density function values, based on the abundance of the different plasmid sizes.

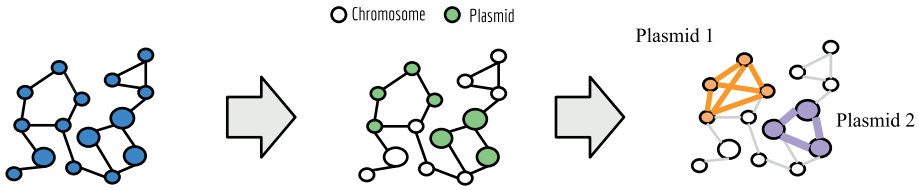


**Supplementary Figure S6.** Number of contigs in which small, large ARG plasmids and large non-ARG plasmids are assembled per species. The 'General' group includes contigs from 66 species less-frequently represented in databases.




Supplementary Figure S7. Recall(contig) values for individual reference plasmids included in benchmark dataset “A”. Plasmids were sub-categorized as small or large with and without resistance genes.

## Individual plasmid reconstruction workflow




**Step 1:** Identification of all plasmid contigs in the assembly graph

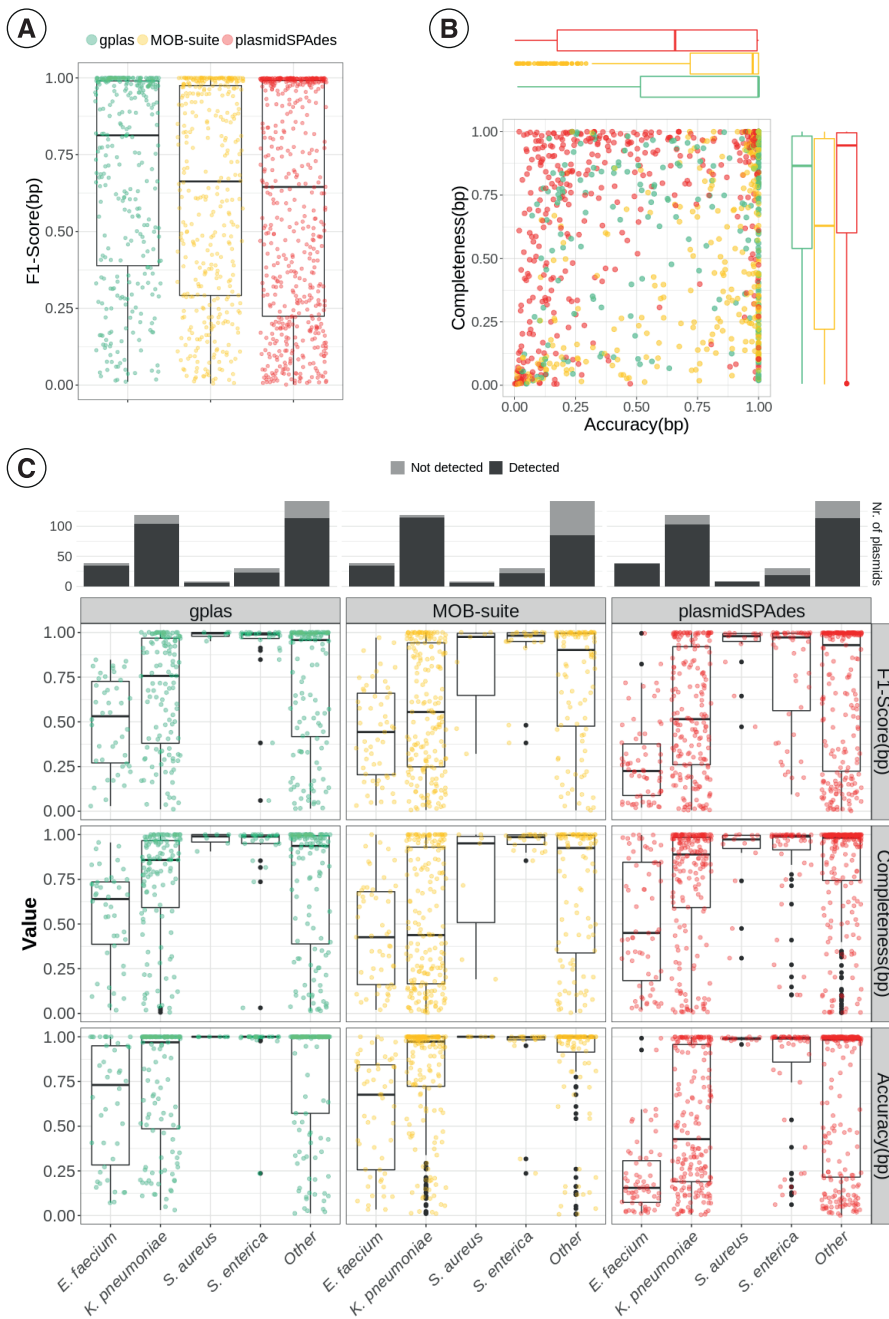
**Step 2:** Binning plasmid contigs into individual predictions


  
**plasmidEC**
  
*E. faecium*, *S. enterica*, *S. aureus*

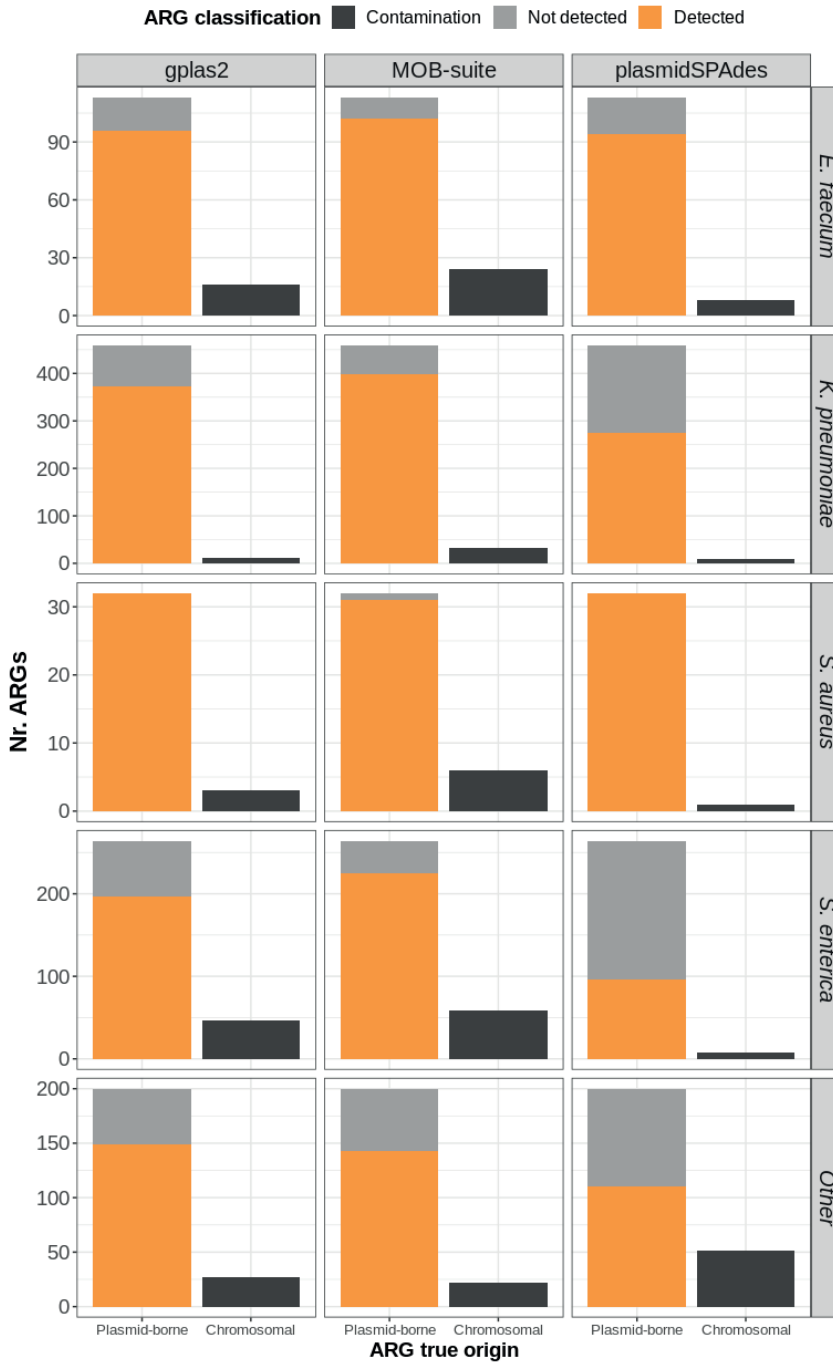
**Centrifuge**
  
*K. pneumoniae*, less-frequent species


  
**gplas**
  
 (For all species)

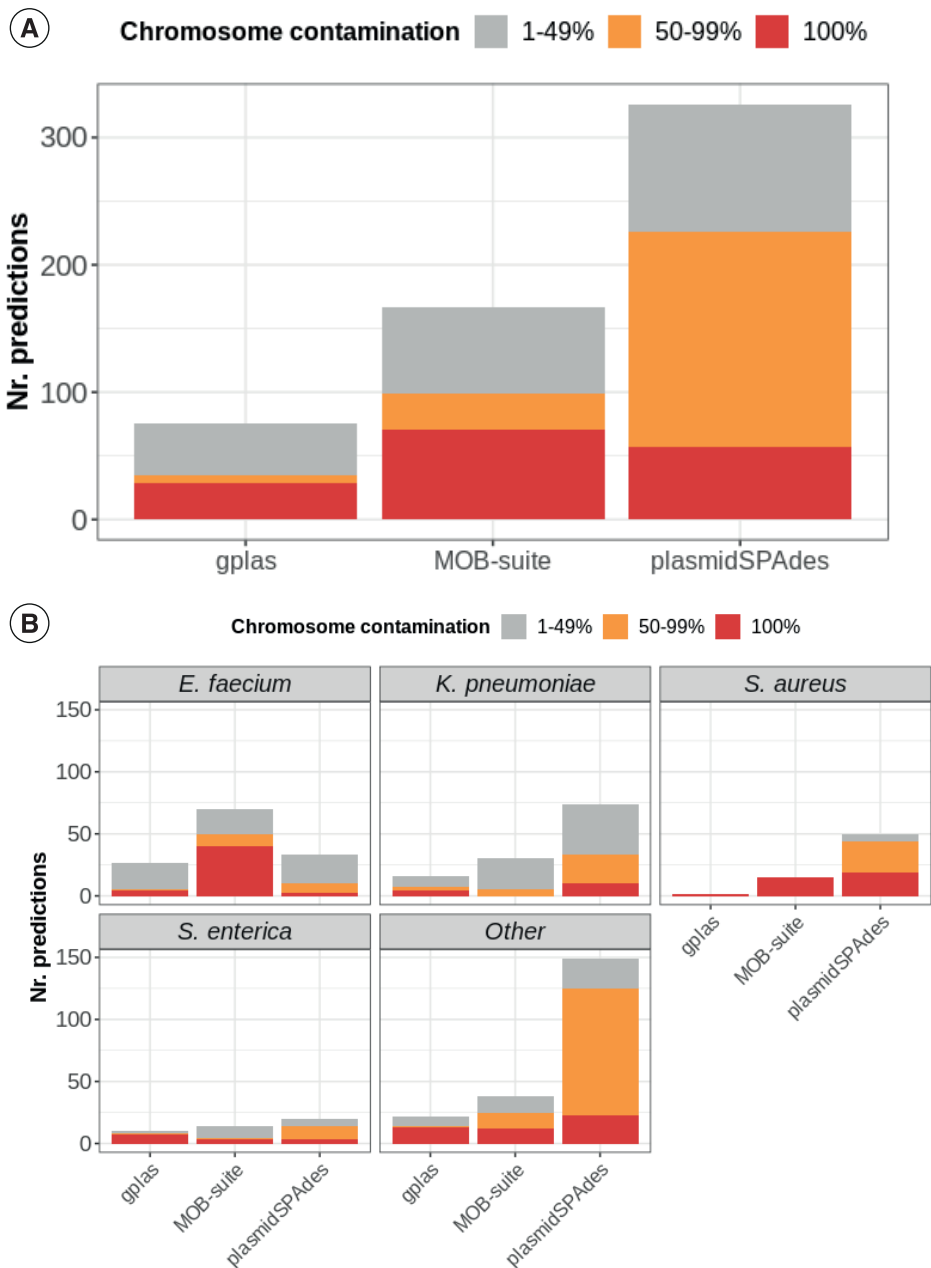
**Supplementary Figure S8.** Schematic representation of the two-step workflow we developed to reconstruct individual plasmids. In step 1, plasmid contigs are identified by using either plasmidEC (*E. faecium*, *S. enterica* and *S. aureus*) or Centrifuge (*K. pneumoniae* and less-frequent species). In step 2, gplas is used to bin plasmid contigs into individual predictions, based on similar sequencing coverage and on node connectivity



**Supplementary Figure S9.** Reconstruction metrics from gplas, MOB-suite and plasmidSPAdes for large plasmids that don't carry ARGs across multiple species (n=338). **A**) F1-Score(bp) and **B**) completeness(bp) vs accuracy(bp) for all plasmids. **C**) Same metrics for each species individually.

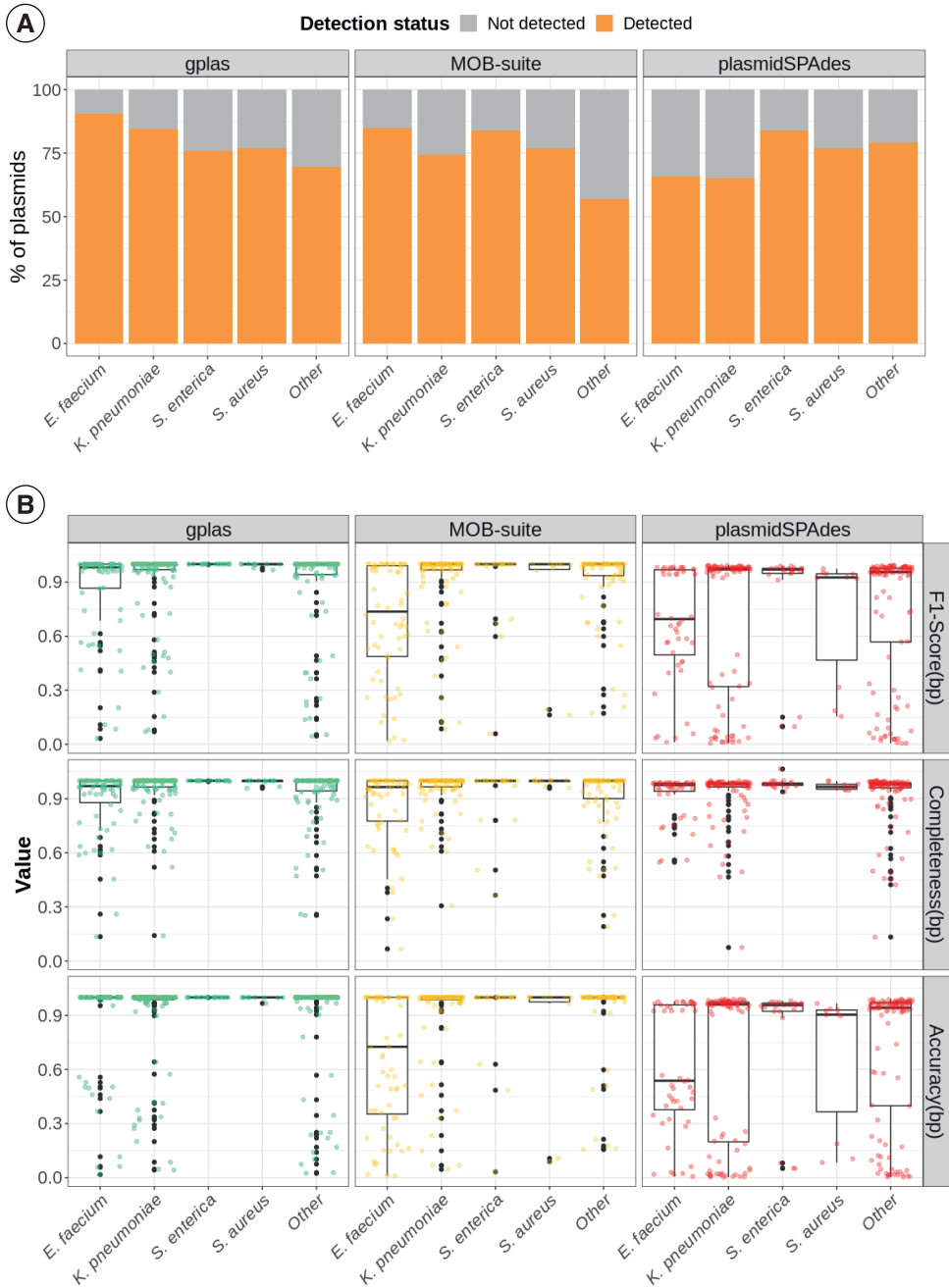


Supplementary Figure S10. Nr. of true plasmid-borne ARGs included (Detected) and missing (Not detected) from plasmid predictions. Chromosomal ARGs included in plasmid predictions are labelled as contamination.

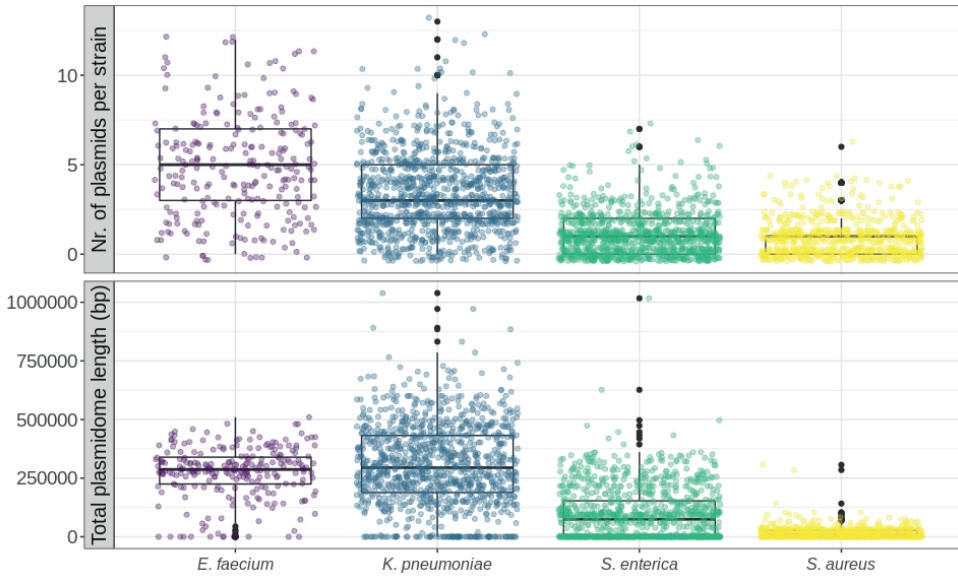


**Supplementary Figure S11.** Nr. of plasmid predictions that include chromosomal sequences (contamination), for all species together in (A) and individually in (B). In red, predictions that are composed solely of chromosomal sequence. In orange, predictions that contain more than 50% of chromosomal sequences in length. In grey, predictions that contain less than 50% of chromosomal sequences.

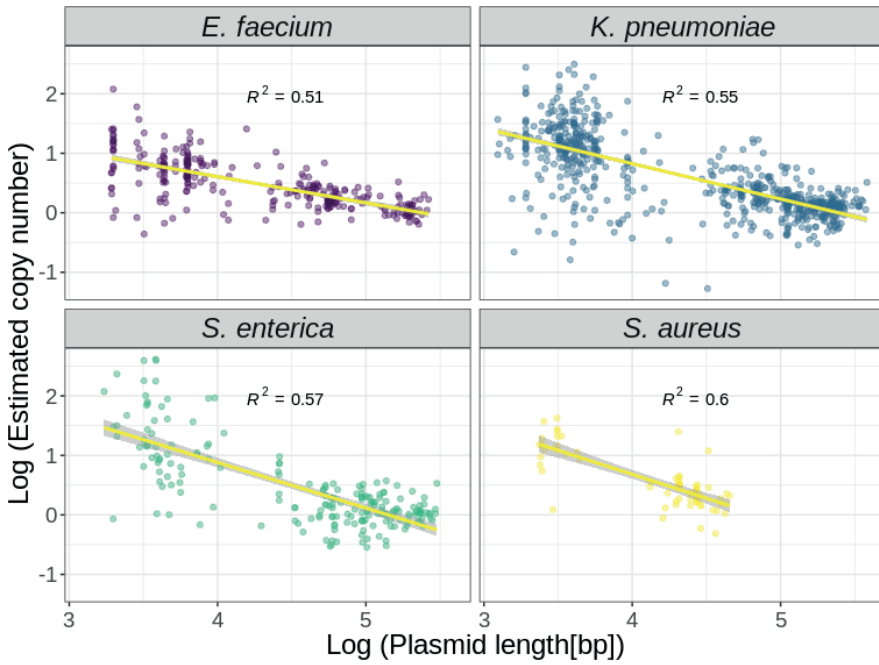




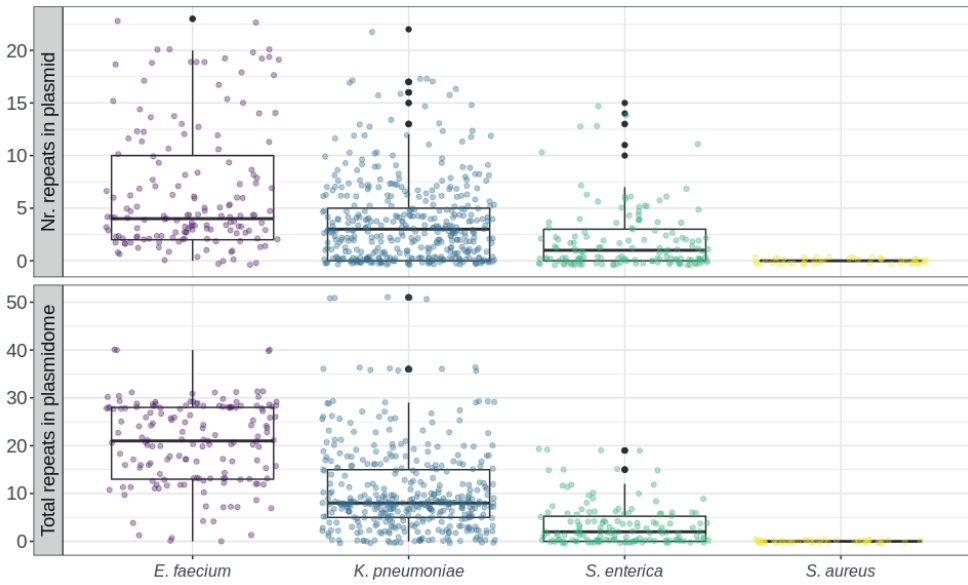
**Supplementary Figure S12.** A) Fraction of small plasmids detected by each tool per species. Small plasmids are defined as those with lengths smaller than 18kb for *E. faecium*, *K. pneumoniae*, *S. enterica* and less-frequent species (Other) and smaller than 8kb for *S. aureus*. B) Reconstruction metrics of small plasmids.



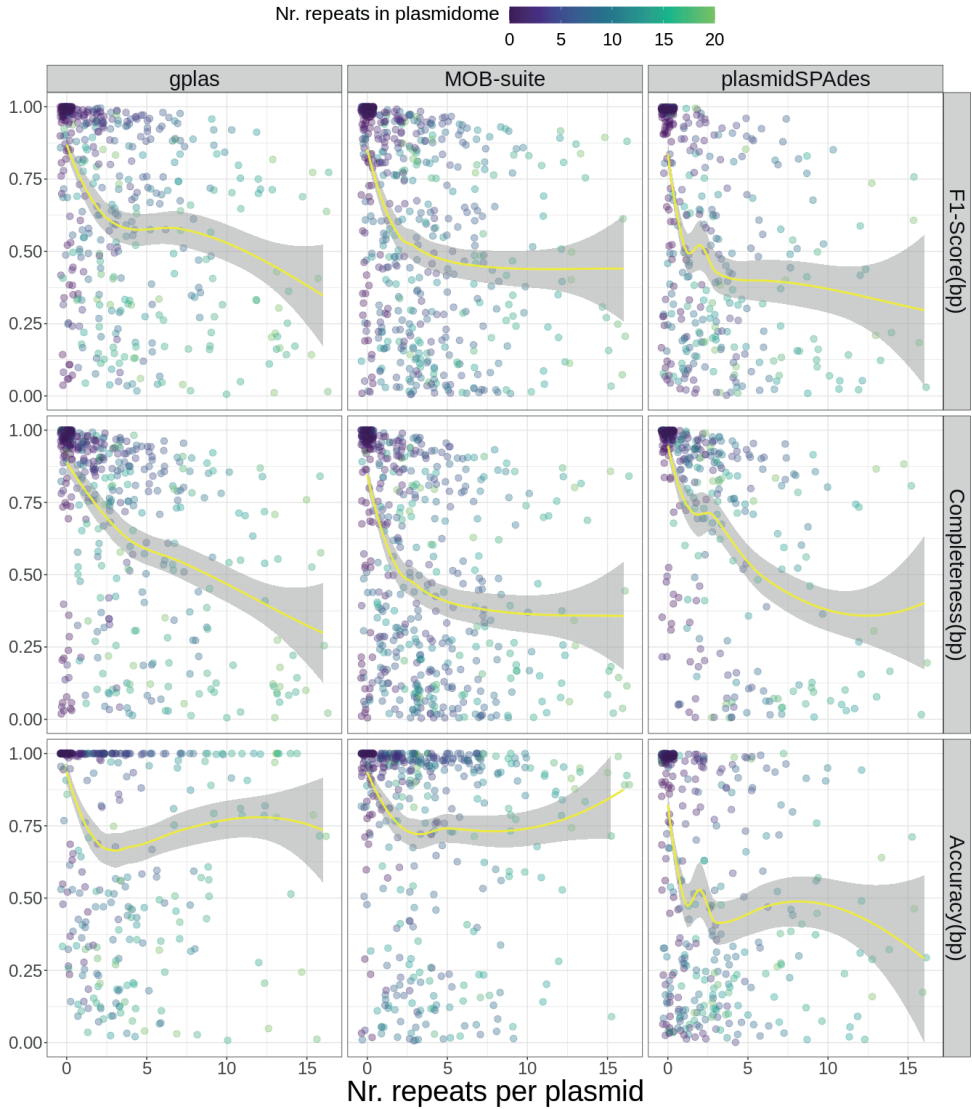
**Supplementary Figure S13.** Number of plasmids per genome (top) and total plasmidome length (bottom) for each species. Results obtained using all complete genomes available for each species in Refseq until July 26th 2021.



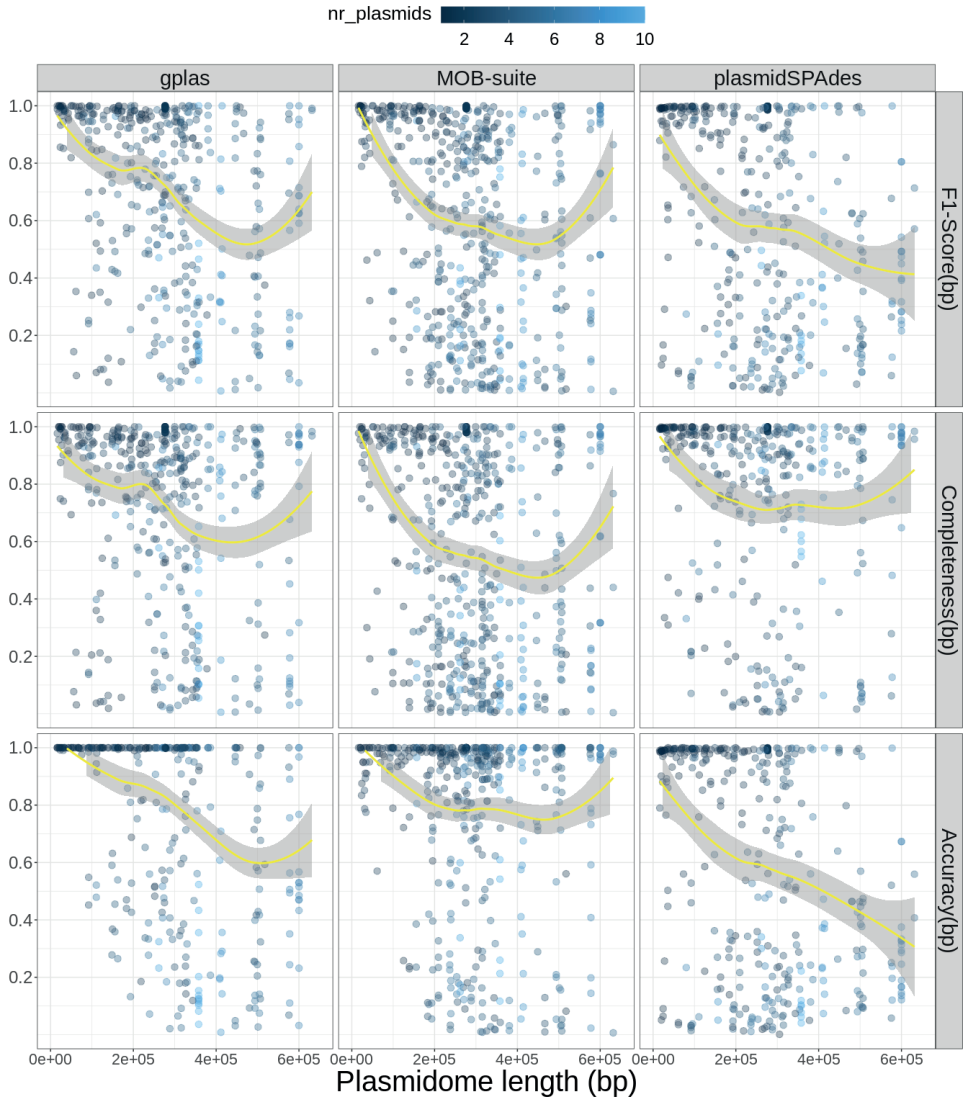
Supplementary Figure S14. Relation between length and estimated copy number for all plasmids included in the benchmarking dataset ( $n=1,923$ ).



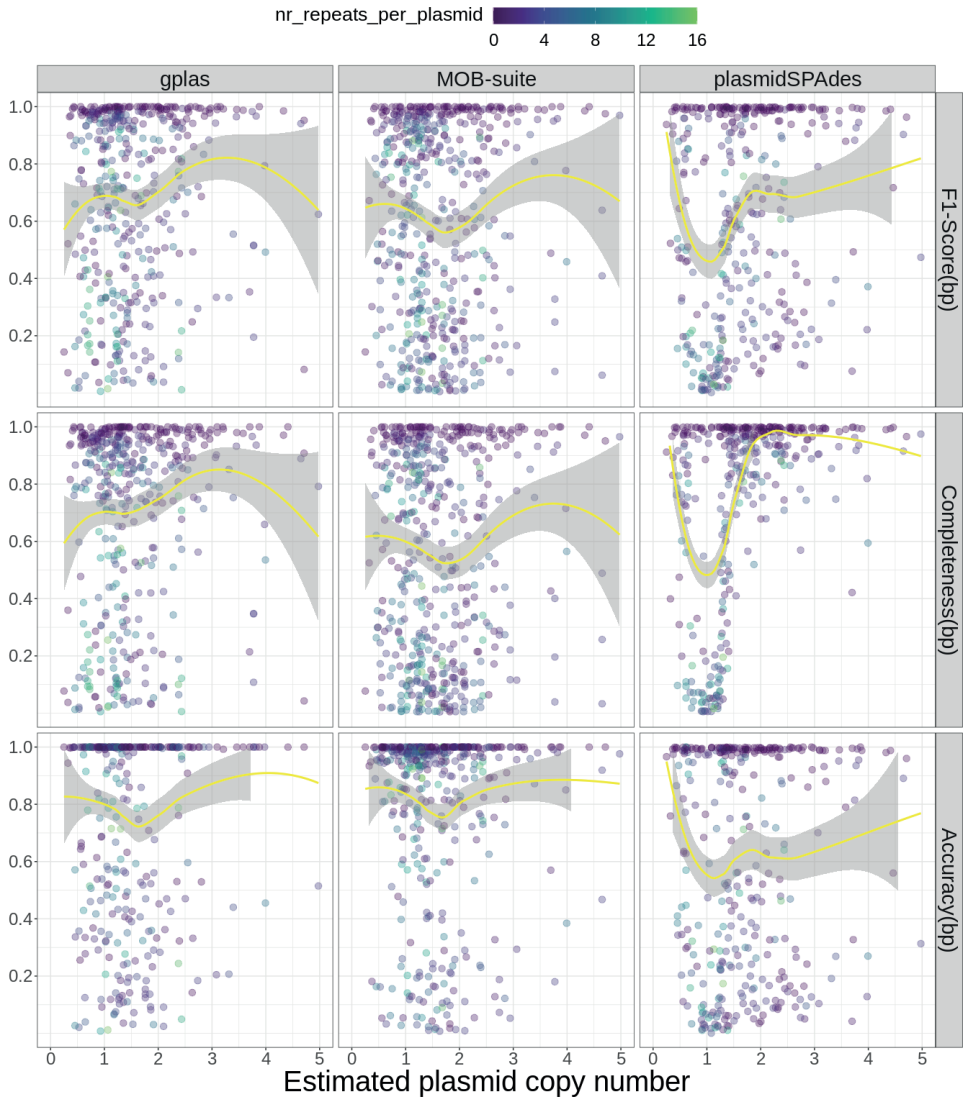
**Supplementary Figure S15.** Number of repeated elements per large plasmid (Top) and total number of repeats in plasmidome per genome (bottom). Results obtained using all genomes included in the benchmark dataset.



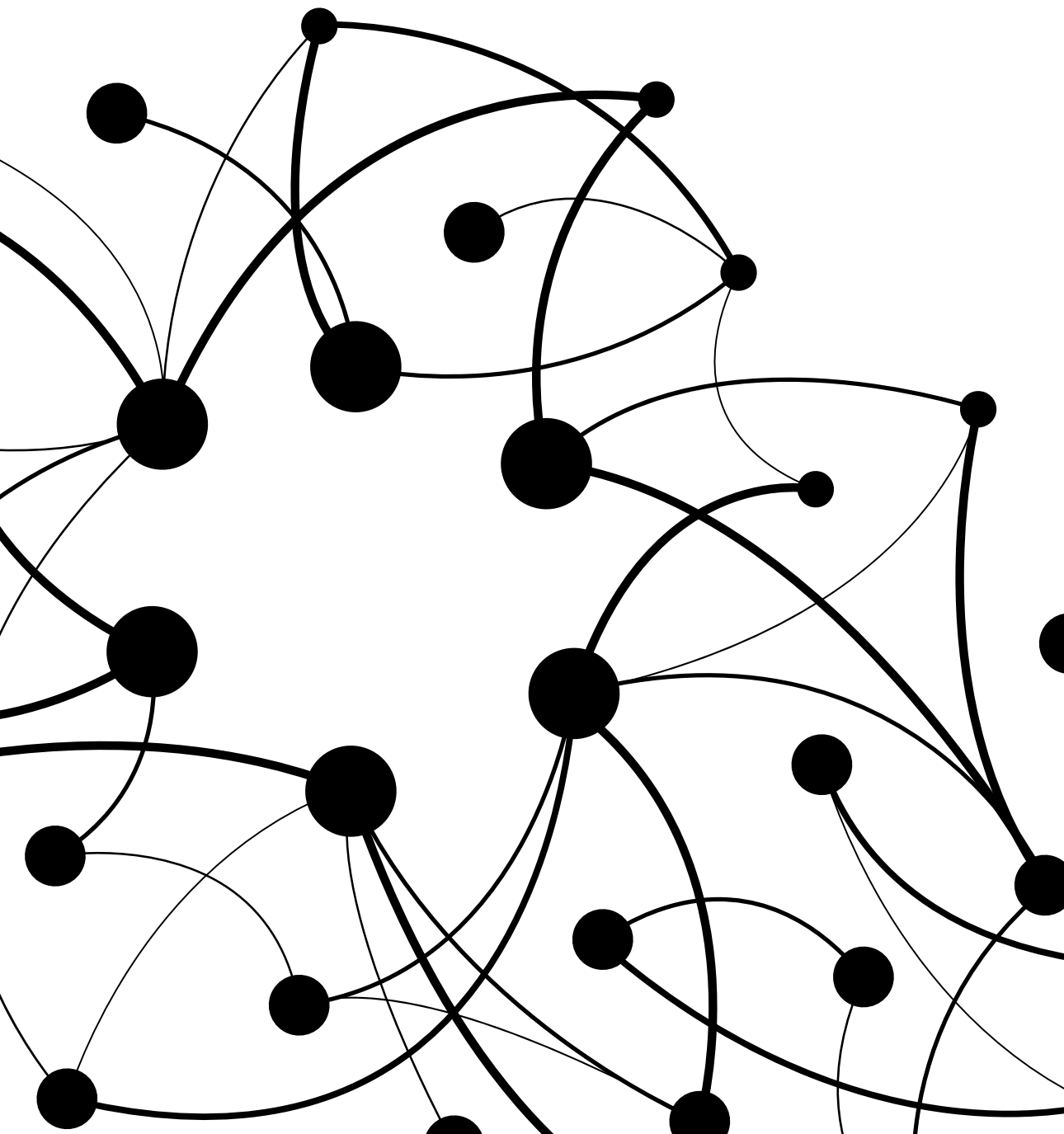
**Supplementary Figure S16.** Reconstruction metrics for large plasmids ( $n=562$ ) plotted as a function of the number of repeats present in each plasmid. The data is colored according to the total number of repeats present in the complete plasmidome of the isolate. Yellow line indicates a LOESS regression ( $\alpha=0.8$ ).



**Supplementary Figure S17.** Reconstruction metrics for large plasmids ( $n=562$ ) plotted as a function of total plasmidome length. The data is colored according to the total number of plasmids present in the isolate. Yellow line indicates a LOESS regression ( $\alpha=0.8$ ).



**Supplementary Figure S18.** Reconstruction metrics for large plasmids ( $n=562$ ) plotted as a function of plasmid estimated copy number. The data is colored according to the total number of repeats in the plasmid. Yellow line indicates a LOESS regression ( $\alpha=0.8$ ).





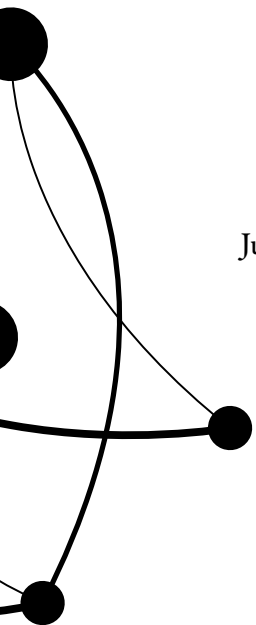
# 05

---

## **Impact of selective digestive decontamination on the pangenome composition of ESBL-*E. coli***

Julian A. Paganini, A. C. Schürch, R. J. L. Willems, N. L. Plantinga  
On behalf of the R-GNOSIS ICU study group

Manuscript in preparation



### Abstract

In Dutch ICUs patients receive selective digestive decontamination (SDD) as a prophylactic antimicrobial treatment to prevent colonisation with potentially pathogenic microorganisms and subsequent infections, with a beneficial effect on 28-day mortality. In the R-GNOSIS ICU study, conducted outside of The Netherlands, SDD consisted of a mix of an oropharyngeal paste and a gastric suspension containing colistin, tobramycin and nystatin. These topical antimicrobial agents aim to target aerobic Gram-negative bacteria, *S. aureus* and yeast. SDD improves patient outcome, but its effects on the resistome and pangenome of potentially pathogenic microorganisms have not been extensively studied. In this work, we compared 129 genomes of *E. coli* isolates from patients that received SDD and patients that did not receive SDD, but standard care only (baseline patients) in five ICUs located across three European countries (R-GNOSIS ICU study). We found that the overall pangenome compositions of *E. coli* recovered from both patient groups were highly similar. Variations in the accessory genome were strongly associated with the phylogeny of isolates but not with the use of SDD. Similarly, the plasmidome variations were not explained by treatment, but rather by the interaction between ICU location and phylogroup. Six antibiotic-resistant genes were significantly more prevalent in baseline patients, and two in SDD. One of these SDD-prevalent genes provides resistance against tobramycin, was flanked by IS26 elements and significantly co-occurred with *bla*<sub>CTX-M-15</sub> in multiple plasmid backbones. Notably, no *mcr* genes coding for colistin resistance were detected.

## Introduction

In the Netherlands, patients admitted to the ICU and undergoing mechanical ventilation receive selective digestive decontamination (SDD) as a prophylactic treatment to prevent colonisation with potentially pathogenic microorganisms (PPMOs). SDD consist of a mix of topical antibiotics (tobramycin, colistin and amphotericin B) targeting aerobic gram-negative bacteria (GNB), *Staphylococcus aureus* and yeast, but leaving the anaerobic flora intact. SDD is administered as an oropharyngeal paste and as a solution through the nasogastric tube. Additionally, a 4-day course of an intravenous cephalosporin (cefotaxime or ceftriaxone) is also provided, to treat any incubating infection at the time of ICU admission [1–3]. A variation of SDD, named Selective Oropharyngeal Decontamination (SOD), consisting only of the oropharyngeal paste, is administered in some ICUs as an alternative to SDD. In the Netherlands, where the prevalence of antibiotic resistance is low [5], SDD was associated with improved patient outcome in comparison to standard care, with reduced mortality, shorter lengths of ICU stay and a lower incidence of ICU-acquired bacteremia [6–10].

The R-GNOSIS ICU study, conducted between 2013 and 2017, compared the effectiveness of SDD, SOD, chlorhexidine 1% mouthwash with standard care alone (baseline) in thirteen ICUs located in six European countries with medium to high prevalence of antibiotic resistance (defined as having an extended-spectrum  $\beta$ -lactamase -ESBL- prevalence of at least 5% of amongst ICU-acquired bacteremia with Enterobacteriaceae) [11]. In this cluster-randomised trial, each treatment was applied during six months in the entire ward (randomized order), to patients with expected length of mechanical ventilation of at least 24h. An important modification of the SDD regime in this study was the absence of the 4-day course of intravenous cephalosporin.

One concern regarding the application of SDD is the exertion of antibiotic pressure in ICUs, which already have the highest levels of antimicrobial use within hospitals [12,13]. In contrast to this, multiple studies suggest that the use of SDD is associated with a decrease in incidence of colonization and infection with antimicrobial resistant microorganisms, both in settings with high and low prevalence of resistance [13 - 15]. On the other hand, a study based on metagenomic data seems to indicate that resistance genes to three classes of antibiotics, namely aminoglycosides, macrolides and tetracyclines, are more abundant in the gastrointestinal tract of SDD treated patients when compared to healthy individuals [16]. Moreover, a separate study concluded that during the application of SDD in a single ICU patient, the burden of two aminoglycoside resistance genes seemed to increase in culturable anaerobic commensal bacteria [17]. Furthermore, these two genes appear to be located on mobile genetic elements (MGEs), increasing the risk of horizontal gene transfer (HGT) from anaerobic bacteria to PPMOs. In addition to the potential selective effects on resistance genes, studies also indicate that SDD treatment alters the gut microbiome composition of ICU patients [18,19]. These ecological changes of the microbiota may also affect the composition of PPMOs, such as *Escherichia coli*, populating the intestinal tract of patients receiving SDD. Therefore, we hypothesise that SDD treatment shapes the pangenome composition of *E. coli*, including the development of resistance.

To address this question, we have sequenced the genomes of 129 *E. coli* isolates from the R-GNOSIS ICU study. These isolates were obtained from patients that received SDD (n=63) or did not receive SDD (n=69, baseline) in five different ICUs located in Spain, Belgium and the UK. We explored the

population structure of these isolates and compared their accessory genome content, plasmidomes and resistomes to determine if SDD leads to the selection of specific genomic features of *E. coli*.

## **Materials and Methods**

### **R-GNOSIS ICU study and selection of isolates for whole genome sequencing**

The R-GNOSIS ICU study was conducted between December 2013 and May 2017 in 13 ICUs from six European countries. A detailed description of the study's aims and methods can be found in [11]. Briefly, to monitor the effect of SDD on colonization with Gram-negative bacteria, surveillance samples were taken twice weekly from the rectum and respiratory tract of all patients included in the study (n=8,496). Samples were inoculated on ESBL selective media (Biomérieux®) and in case of growth, phenotypic susceptibility testing was performed according to local standard operating procedures (for colistin susceptibility testing, E-tests were provided) [11]. Clinical blood and respiratory samples were obtained at discretion of the clinician and processed according to local laboratory protocols. From these cultures, unique highly-resistant microorganisms were stored for whole genome sequencing (WGS) according to the following rule: one isolate per patient, per body site, per species, with a unique phenotypic resistance pattern.

We submitted for WGS all stored *E. coli* isolates from the SDD and baseline periods of the five hospitals with most stored isolates (AN, PS, UZ, LB and CD). Isolates were only included if they occurred from day 2 of inclusion onwards (with the date of study enrollment being day 0), to ensure sufficient exposure to the antimicrobials used in SDD. Metadata associated with sequenced isolates can be found in Supplementary Data 1.

### **Whole Genome Sequencing**

Selected isolates were sequenced using Illumina MiSeq, with a Nextera XT pair-end kit (2 x 150bp). Short reads were quality trimmed with trim-galore (v0.6.6) (<https://github.com/FelixKrueger/TrimGalore>). Assembly of genomes was performed with Unicycler (v0.4.9) [20].

### **Pangenome analysis**

Genomes were first annotated with BAKTA(v1.6.1) [21]. Panaroo [22] was then used to define core and accessory genes. A core gene was defined as being present in 99% of all sequenced isolates.

Using the presence/absence gene matrix generated by Panaroo, we calculated Jaccard distances between accessory genomes of all pairs of isolates as:

$$JaccardDistance = 1 - \frac{Acc\ genes\ of\ Insolate\ 1 \cap Acc\ genes\ of\ Insolate\ 2}{Acc\ genes\ of\ Insolate\ 1 \cup Acc\ genes\ of\ Insolate\ 2}$$

Pangenome accumulation curves were obtained using the R micropan package [23].

### Population structure determination

Multi-locus sequence types of isolates were predicted *in silico* with mlst (v2.1.6) (<https://github.com/tseemann/mlst>). Phylogroups were predicted with ClermonTyper (v20.03) (<https://github.com/A-BN/ClermonTyping>). PopPUNK (v2.4) [24] was used to assign draft genomes to existing clusters according to the *E. coli* database (v1) available at (<https://www.bacpop.org/poppunk/>).

A Neighbour-joining tree was constructed using IQ-TREE (v2.2.0.3), based on a core-genome alignment obtained with Panaroo.

### Plasmidome analysis and plasmid reconstructions

The plasmidome of each genome was defined as all plasmid-predicted contigs identified by using plasmidEC (v1.3) (<https://gitlab.com/mmb-umcu/plasmidEC>). Similar to previously described, plasmidomes were annotated with BAKTA, and Jaccard distances between these were calculated based on the presence/absence gene matrix generated by Panaroo.

Individual plasmids were reconstructed using gplas (v1.1) [25]. Distances between all plasmid predictions were obtained using MASH (v2.2.2) [25,26] with k-mer length of 21, and a sketch size of 10,000. Clusters of highly similar plasmids were obtained by creating a network in which connections between plasmids were drawn if their MASH distance was below 0.01.

Clusters of plasmids backbones were created using mge-cluster(v1.1) [27] and clusters numbers were assigned based on the existing *E. coli* database, which can be accessed at: <https://doi.org/10.6084/m9.figshare.21674078.v1>.

### Comparison of plasmidome and accessory genomes

Distances between accessory genomes and plasmidomes of isolates were visualised using t-distributed stochastic neighbour embedding algorithm (t-SNE), as implemented in the Rtsne R package (v0.15).

To conduct permutational analysis of variance (PERMANOVA), we used the adonis function from the vegan R package (v2.5-6) using the matrix of pairwise Jaccard distances as input. To explain the variance of accessory genome and plasmidome distances, six different PERMANOVA models were built with different explanatory variables each, as detailed in Supplementary Tables S1 and S2. In models with two variables, interaction terms between them are indicated with “\*”.

### Estimation of plasmid copy number

After short-read assembly with unicycler, each contig is assigned a relative coverage value. We used all unitigs that unambiguously aligned to a single replicon to calculate the mean relative coverage of each plasmid. Duplicated contigs, aligning to more than one location of the genome, were left out of these calculations.

**ARGs, co-occurrence networks, and genomic context**

Antibiotic resistance genes were identified by using AMRFinderPus (v3.11.2). Co-occurrence of ARGs in the same plasmids were calculated by using a previously described approach [28]. The genomic context of the tobramycin resistance transposon and of *bla*<sub>CTX-M-15</sub> genes were obtained by manually exploring the assembly graph, and using BLAST against the ISFinder database of the nodes that surrounded the aforementioned elements.

**Statistics and code availability**

Comparison of medians and proportions was performed using the non-parametric test Wilcoxon rank sum test [29] and Fisher's exact test [30], respectively. Statistical analysis was performed using R (v3.6.1) and code needed to reproduce this analysis can be found in: [https://gitlab.com/jpaganini/rgnosis\\_sdd\\_baseline](https://gitlab.com/jpaganini/rgnosis_sdd_baseline).

**Results****Patients colonised with ESBL-*E. coli* in five European ICUs**

The five selected ICUs (AN, PS, UZ, LB and CD) were located in five different hospitals in Belgium (n=3), Spain (n=1) and the UK (n=1). In total, 129 isolates were obtained from 116 patients, and in most cases (n=103, 90.6%), a single isolate per patient was sequenced. Most sequenced isolates derived from the ICU termed LB (n=55), while the rest of isolates were similarly distributed across the remaining locations ranging from 16 (PS) to 21 (AN) isolates per ICU (Table 1). There was a similar number of isolates from SDD (n=63) and baseline (n=66) periods. The majority of isolates (n=124, 96.1%) were obtained from rectal swabs, while a small number derived from respiratory samples (n=4, 3.1%) and bloodstream infections (n=1, 0.8%). Also, the majority of isolates (n=122, 94.6%) were phenotypically resistant to cefotaxime, ceftriaxone and/or ceftazidime. Although not all isolates have phenotypic ESBL confirmation, in this manuscript, we refer to the isolates as ESBL-*E. coli*. Sequenced samples were obtained after a median of 4 days (IQR= 2 - 6.5) after the start of SDD vs. 6 days (IQR=4 - 11) after inclusion in the baseline period. All metadata associated with patients and sequenced isolates can be found in Supplementary Data 1.

**Population structure of colonising ESBL-*E. coli***

Sequenced isolates belonged to 54 different STs, 11 of these were present in both study periods, 24 STs were only found in isolates from the baseline period and 19 were exclusively found from SDD patients (Supplementary data 1). ST131 was the predominant clone (n=30) in both groups (baseline n=16, 24%; SDD n=14, 22%) (Figure 1A), followed by ST410 (n=7), ST10 (n=6) and ST1193 (n=5). We used PopPUNK to assign isolates to existing clusters considering both core and accessory genome variations (Supplementary data 1). We found a total of 53 clusters, of which the most abundant was Cluster 2 (n=27, 21%), which was entirely composed of ST131 isolates. Cluster 7\_510, the second most abundant (n=12, 9.3%), encompasses isolates from ST10 (n=4), ST167 (n=3), ST744 (n=4) and ST1695 (n=1). A more detailed exploration of the core-genome of isolates (Figure 1B) showed no clear clusters associated with treatment, suggesting that SDD does not select for particular *E. coli* clones or lineages.

**Table 1.** Sequenced isolates according to study period, hospital, body site and time of isolation

			Baseline	SDD	Total
Nr. of sequenced isolates	Per hospital	LB	28	27	55
		CD	13	7	20
		PS	5	11	16
		AN	14	7	21
		UZ	6	11	17
	Per body site	blood	1	0	1
		rectum	62	62	124
		respiratory	3	1	4
Median nr. of days between inclusion and culture collection of the sequenced isolate (IQR)		6 (4 - 11)	4 (2 - 6.5)		

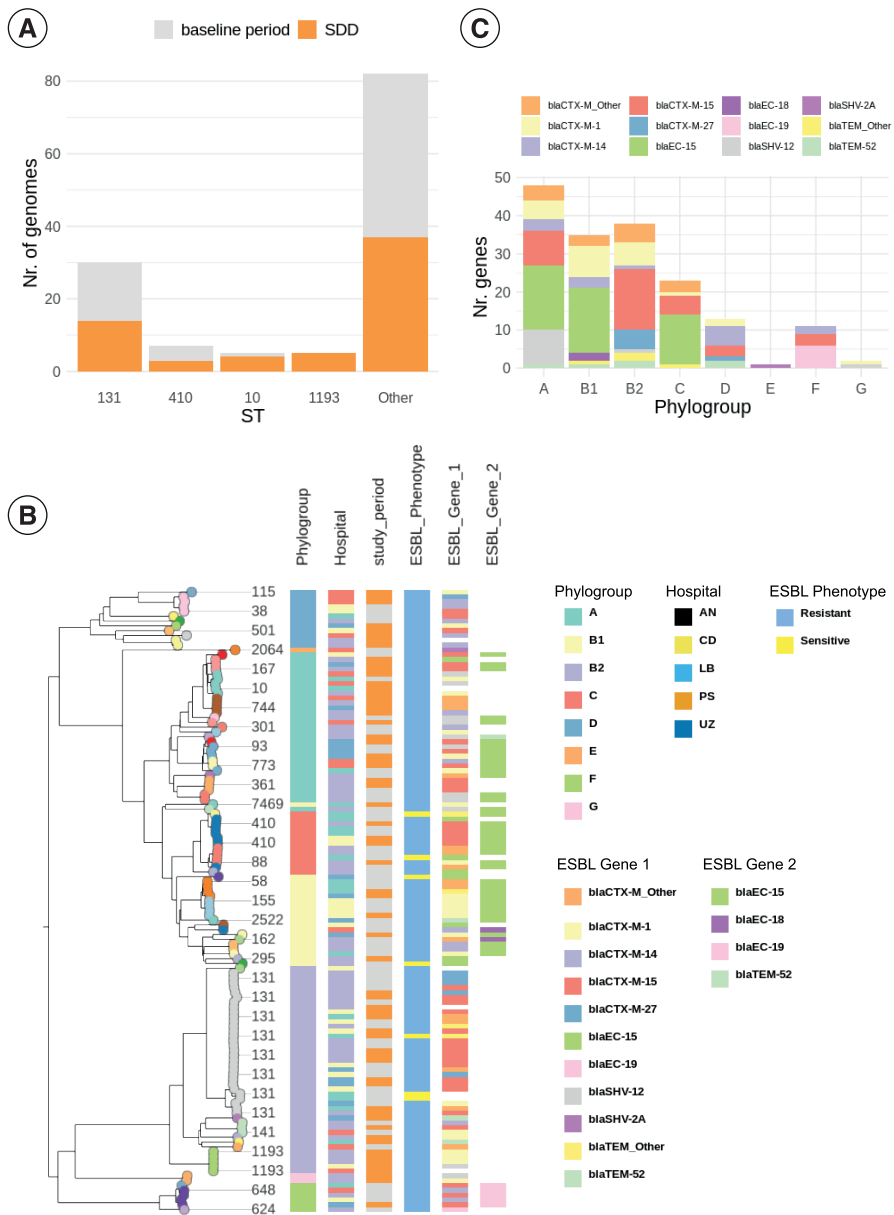
Simpson's indices of diversity calculated using ST (baseline=0.919, SDD=0.921) and PopPUNK clusters (baseline=0.924, SDD=0.918) indicated that the population structure in both study periods was equally diverse.

Predictions of ARGs from our dataset showed that most isolates (n=123, 95.3%) carried at least one ESBL gene (Figure 1B). All isolates classified as phylogroup B2 (n=42) carried only one ESBL gene, with *bla*<sub>CTX-M-15</sub> being the most abundant allele (n=16, 38.1%), followed by *bla*<sub>CTX-M-1</sub> (n=6, 14.3%) and *bla*<sub>CTX-M-27</sub> (n=5, 11.9%) (Figure 1B and C). Similarly, samples belonging to phylogroup D (n=12) also carried one ESBL gene, but *bla*<sub>CTX-M-14</sub> was the most prevalent variant (n=5, 41.6%). In contrast, the majority of isolates from phylogroups A (17/31), B1 (15/20) and C (10/13) carried two ESBL genes, and *bla*<sub>EC-15</sub> was the most frequent gene in all phylogroups, with prevalences of 35%, 48% and 57%, per phylogroup respectively.

### Variations in accessory genome compositions are associated with phylogroup and hospital, but not with the use of SDD

Considering all 129 isolates, the total pangenome consisted of 13,753 genes, of which 3,072 were identified as core- and the remaining 10,681 as accessory-genes. Accumulation curves fitted to Heap's law for baseline and SDD isolates yielded similar alpha values (baseline=0.87, SDD=0.91) (Supplementary Figure S1A), confirming an open pangenome for both groups[31]. The number of accessory genes in isolates from the baseline (median=1,645; IQR=1,503 - 1,776) and SDD periods (median=1,642; IQR=1,476-1,775) did not differ (p-value=0.92, Wilcoxon rank sum test) (Supplementary Figure S1B).

To identify associations between gene frequency and function, we obtained the functional categories of Clusters of Orthologous Groups (COG) from BAKTA annotations. A COG function was assigned to 66.1% of all predicted coding sequences (CDS). We performed a Fisher's exact test to compare the frequency of each COG category in baseline and SDD isolates. This analysis showed that the distribution of genes to COG categories was similar between isolates from both study periods (Supplementary Figure S1C, Supplementary Table S3).



**Figure 1.** A) Distribution of most abundant ST across study periods. Groups with less than 5 isolates were collapsed into the ‘Other’ category. B) Neighbor-joining phylogenetic tree constructed based on core-genome alignment. Labels on the leaves indicate ST of isolates. Phylogroups were predicted *in silico* using ClermonTyper. ESBL genes were predicted with AMR-FinderPlus, if a second ESBL gene was present in an isolate, this is indicated in the column ESBL\_Gene\_2. ESBL phenotype indicates phenotypic resistance or sensitivity to third-generation cephalosporins, evaluated as indicated in Methods. C) Distribution of the different ESBL genes across phylogroups



Next, we explored the diversity in total accessory gene content of individual isolates and its association with phylogroup, hospital and treatment by using PERMANOVA (Figure 2A and 2B, Supplementary Table S1). The accessory genome composition was strongly associated with the isolate's phylogroup ( $R^2=0.439$ ,  $p$ -value=0.001), and its interaction with the geographical location ( $R^2=0.10$ ,  $p$ -value=0.006). There was no significant effect of SDD in the accessory genome composition ( $R^2=0.01$ ,  $p$ -value=0.149).

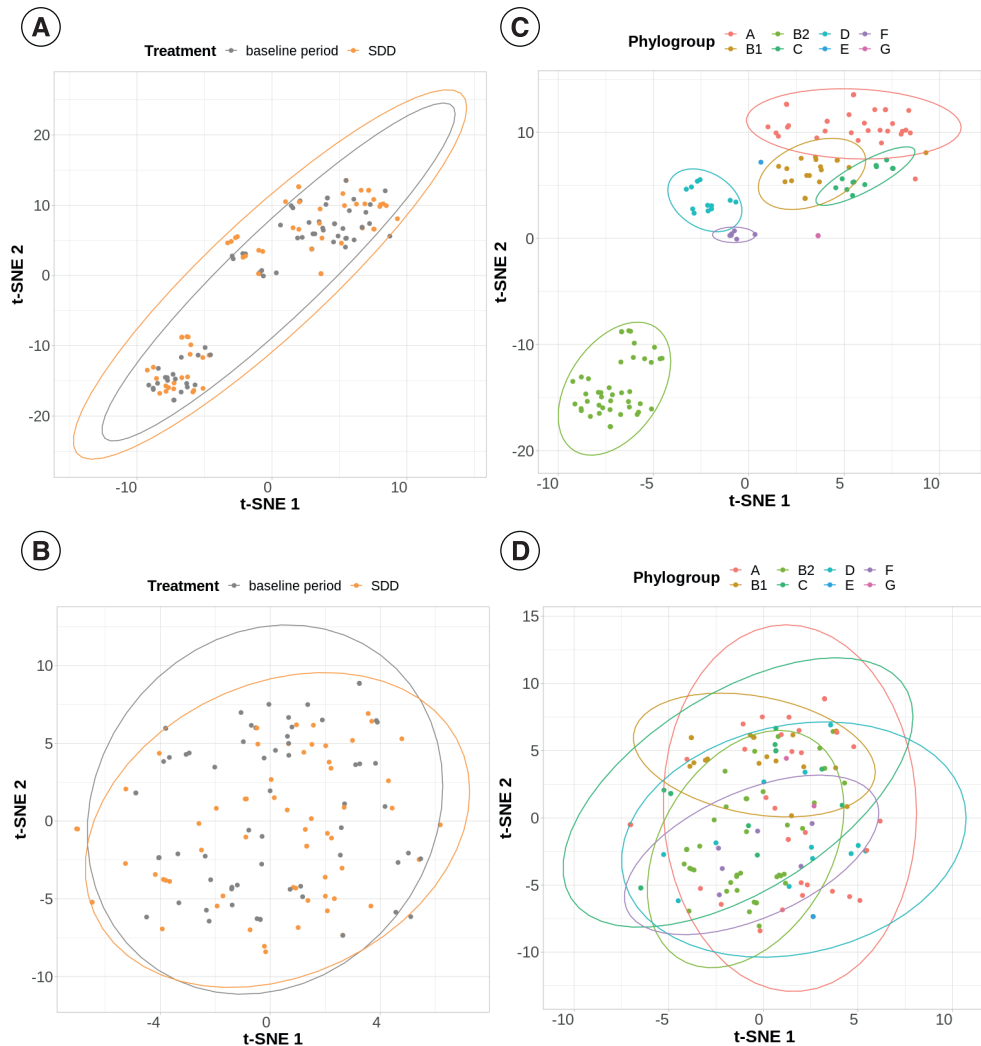
The plasmidome is important for niche adaptation in *E. coli* [32] and other gut bacteria [33]. Consequently, we also compared predicted plasmidome compositions across study periods. The median of plasmidome sizes for baseline isolates was 196,520.5 bp (IQR= 144,476.5 - 277,834.5) and of 216,196.0 bp (IQR= 141,770.0 - 293,638.8 ) for SDD isolates, which was not statistically different ( $p$ -value=0.92, Wilcoxon rank sum test, Supplementary Figure S2A). Additionally, the median number of unique plasmids per isolate, as predicted by gplas reconstructions, was 4 in isolates from both study periods ( $p$ -value=0.92, Wilcoxon rank sum test, Supplementary Figure S2B). Sizes and copy number of individual plasmids also followed expected distributions in both study periods [32,34] (Supplementary Figure S2C).

The use of SDD was also not associated with plasmidome variation ( $R^2=0.011$ ,  $p$ -value=0.075, PERMANOVA), but the interaction between phylogroup and the ICU explained the largest fraction of plasmidome variation ( $R^2=0.154$ ,  $p$ -value=0.002, PERMANOVA) (Figure 2C and 2D, Supplementary Table S2).

Since the ICU had a significant effect on the plasmidome composition, we wanted to evaluate if this was related to the fact that highly specific plasmid (or clones) were persistent over time in each ICU. For this, we predicted individual plasmids using gplas, and clustered these plasmid predictions based on MASH distances (Supplementary Figure S3A and S3B). A total of 558 plasmids were predicted in 128 isolates. Of these, 257 plasmids (46%) were grouped into 65 clusters composed of highly similar plasmids (MASH distance < 0.01). Interestingly, 37% of these clusters were exclusively composed of plasmids isolated from a single hospital, while 63% included plasmids from multiple hospitals. Moreover, 41.7% ( $n=10/24$ ) of plasmid clusters recovered in single hospitals were found associated with multiple clones (PopPUNK clusters) (Supplementary Figure S3C, Supplementary data 2).

### **A putative mobile genomic element encoding a tobramycin resistance gene is enriched in isolates from SDD treated patients**

A total of 100 unique ARGs were found in the entire dataset (Supplementary data 3). Isolates obtained during baseline treatment contained a similar number of ARGs (median=12, IQR = 8 - 14) as those of SDD (median=11, IQR= 7 - 13) ( $p$ -value=0.27, Wilcoxon ranked-sum test) (Supplementary figure S4). Surprisingly, when subclassifying ARG by antibiotic class, we found a higher number of aminoglycoside resistance genes in baseline isolates (median= 3, IQR=1.25 - 4) than in SDD isolates (median=2, IQR= 1 - 3) ( $p$ -value=0.03, Wilcoxon ranked-sum test). For other ARG classes, no significant differences across study periods were found.



**Figure 2.** t-SNE plots representing Jaccard distances between complete accessory genomes (A and B) and predicted plasmidomes (C and D) for 129 ESBL-*E. coli* genomes included in this study. Jaccard distances were calculated using gene presence/absence.

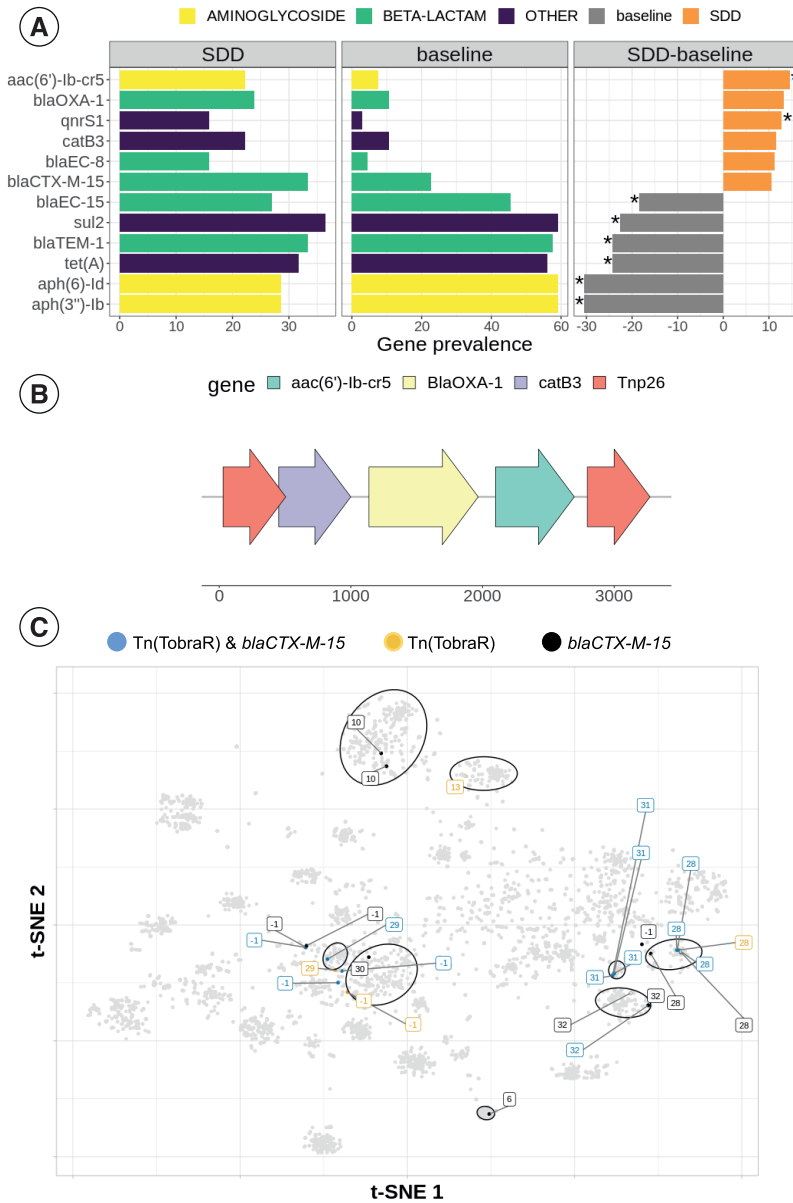
We then compared the occurrence of individual ARGs (Supplementary Figure S5) in both study periods. A total of 12 ARGs had an absolute difference in prevalence larger than 10% across study periods (Figure 3A). Out of these, six genes were significantly more frequent in baseline isolates, including two that code for beta-lactamases (*bla*<sub>EC-15</sub>, *bla*<sub>TEM-1</sub>), two genes that provide resistance to streptomycin [*aph*(6)-*ld*, *aph*(3'')-*lb*], and two ARGs that provide resistance to sulfonamide (*sul2*) and tetracyclin (*tet*(A)). Additionally, two genes were significantly more frequent during SDD, namely *aac*(6')-*Ib-cr5* and *qnrS1* (Fisher's exact test).

The *aac(6′)-Ib-cr5* gene encodes an aminoglycoside 6′-N-acetyltransferase, which is predicted to provide resistance to tobramycin, one of the components of SDD [1]. It was found in 5/66 (7.5%) isolates during baseline and in 14/63 isolates (22.2%) during SDD. In the majority of SDD isolates (n=13/14, 92.9%), *aac(6′)-Ib-cr5* was encoded in conjunction with *bla<sub>OXA-1</sub>* and with *catB3* on a ~2.2Kb contig (Figure 3B), which was flanked by IS26 elements (Supplementary Figure S6). This suggests that this element is a potentially mobile composite transposon, termed from here on Tn(TobraR).

To evaluate if Tn(TobraR) was found in different plasmid backbones, we used mge-cluster to assign the plasmid predictions generated by gplas to existing *E. coli* plasmid clusters (see methods). We found Tn(TobraR) in five distinct plasmid backbones (Figure 3C), namely cluster 13 (n=1), 28 (n=4), 29 (n=2), 31 (n=4) and 32 (n=1), and also in 5 different plasmids that were not assigned to a previously existing plasmid type, supporting the hypothesis that Tn(TobraR) can actually move independently. Additionally, this element was found in isolates from multiple chromosomal backgrounds in our dataset, including phylogroups B2 (n=12, most ST131), C (n=1), F (n=2) and A (n=2) (Supplementary Figure S7A). Moreover, when querying a database composed of more than 1,300 publicly available *E. coli* genomes, we found the Tn(TobraR) in 63 additional genomes from six different phylogroups (Supplementary Figure S7B). Notably, in SDD isolates, we observed that the ARGs that compose the Tn(TobraR) co-occurred with *bla<sub>CTX-M-15</sub>* in the same plasmid significantly more frequently than expected by chance (Supplementary Figure S8A, Supplementary table S4). However, this was not the case in baseline isolates (Supplementary Figure S8B, Supplementary table S5).

Following the statistical analysis on pre-selected genes which had a 10% difference in prevalence between baseline and SDD isolates, we also compared the prevalence of all tobramycin resistance genes. There were no significant differences in the occurrence of other tobramycin resistance genes other than *aac(6′)-Ib-cr5* (Supplementary Figure S8C).

Finally, we evaluated the prevalence of genes coding for carbapenem resistance. A total of four isolates were predicted to have a carbapenemase gene, all of which were obtained from SDD-treated patients. One of these isolates carried a *bla<sub>OXA-48</sub>* gene in an ST295 background; two isolates from the same patient, belonging to ST410, coded a *bla<sub>OXA-181</sub>* gene; and *bla<sub>VIM-1</sub>* was found in an ST1193 isolate.



**Figure 3.** A) Prevalence of 12 antibiotic resistance genes (ARGs) in isolates from baseline (left) vs SDD (center) patients. The difference in prevalence between the two study periods is displayed on the most right panel. ARGs are displayed only if an absolute difference in prevalence greater than 10% was observed across study periods. Fisher’s exact test was used to determine which genes were significantly more prevalent across study periods (\*). B) Putative transposon carrying the *aac(6)-Ib-cr5* gene, termed Tn(TobraR). C) t-SNE plot in which each dot represents an *E. coli* plasmid. Grey dots represent complete plasmids obtained from NCBI database ( $n \sim 4,500$ ). Coloured dots represent plasmid predictions of isolates from patients that received SDD treatment and that carry Tn(TobraR), *bla*<sub>CTX-M-15</sub> or both. Labels and ellipses depict different plasmid types, obtained with mge-cluster (v1.1). Labels equal to [-1] correspond to plasmid not assigned to any plasmid type.

## Discussion

In this work we sequenced the genomes of 129 ESBL-*E. coli* isolates of which 63 were obtained from ICU patients that received SDD as a prophylactic measure to prevent colonisation and infection with potentially pathogenic gram-negative bacteria, and compared those with 66 genomes of ESBL isolates from the same patient population that did not receive SDD. One of the main strengths of our study is its multi-center nature, with isolates from five different ICUs located in three European countries. Moreover, in contrast to many similar studies [35–39], 96% of the isolates represented intestinal carriage isolates, rather than being recovered from clinical samples.

The results from our study revealed no important differences in pangenome compositions between ESBL-*E. coli* recovered from patients treated with SDD and those not treated with SDD. This suggests limited impact of SDD in shaping the overall pangenome composition of ESBL-*E. coli*. These results were unexpected when considering that SDD alters the microbial composition of the gut [18,19], potentially changing the interaction networks that occur in the microbiota leading to new metabolic challenges for *E. coli* [40].

The absence of important differences in the pangenome can have several potential explanations. First, it is possible that the adaptation of ESBL-*E. coli* to the new gut ecology induced by SDD is not mediated by the acquisition/loss of certain genes, but rather by changes in gene expression patterns. These changes cannot be detected by our analysis which solely relies on gene content comparisons. Supporting this hypothesis, a recent study has described that a global re-wiring of transcription factors is observed when switching *E. coli* from auxotrophic to prototrophic growing conditions [41]. Moreover, a recent study demonstrated that *E. coli* auxotrophies can be rescued by expressing short peptides that are coded in novel small open reading frames, which will also be missed by the annotation tools that we have used in this study. Second, it is also possible that the duration of SDD (with a median time of isolation of the first ESBL-*E. coli* in SDD being 4 days) was not sufficient to cause an impact in the community structure of *E. coli*. Finally, the fact that we had only ESBL-positive isolates available for WGS, prohibited analysis of changes in the relative abundances of different *E. coli* subpopulations within each patient. Future research should ideally collect multiple isolates (or fecal samples) over longer periods of time from the same patients, including also non-ESBL *E. coli*.

The interplay between phylogeny and the ICU explained the largest fraction of variance observed in the accessory genome and plasmidome of isolates. The strong association between phylogeny and the accessory genome of *E. coli* has already been described in [32,43]. The effect of the ICU could be explained by postulating that each ward constitutes its own ecosystem, in which particular plasmids, and clones, persist over time, with the ability to spread to different patients. In line with this hypothesis, recent studies suggest that plasmids carrying carbapenem resistance genes present ‘geographical signatures’ that relate them to particular healthcare settings [44]. Moreover, the long-term persistence of clones and plasmids in clinical environments is more common than originally thought [45–49]. A patient admitted to an ICU can be colonised by bacteria contaminating this environment in less than 6 days [45]. On the broader scale, particular plasmids have also been associated with specific countries [50,51].

We also observed minor differences between the resistomes of SDD and baseline isolates. While six different resistance genes were more prevalent in baseline isolates, SDD isolates were enriched in a potentially mobile genetic element composed of a tobramycin resistance gene and two additional ARGs, flanked by IS26 elements. An identical genetic element was reported in other studies [52,53], but its mobility as an independent unit was never tested. It is well known that IS26 plays a crucial role in the dissemination of resistance genes among Enterobacteriaceae in the clinical environment [54–56]. IS26 catalyses a highly efficient conservative transposition reaction that allows the incorporation of ARGs preferably into replicons that contain a pre-existing copy of this element [57,58]. This mechanism could lead to the formation of arrays of in-tandem resistance genes, also referred to as resistance islands [54]. In line with this, we observed that the tobramycin resistant transposon identified in this work frequently co-occurred with a *bla*<sub>CTX-M-15</sub> gene in multiple plasmid backbones that reside in distinct *E. coli* clones. This means that if SDD in fact selects for this transposon, it could also facilitate the formation of resistance islands that accommodate multiple ARGs. However, it is important to note that we had insufficient isolates to perform multivariable analysis that can correct for population structure, such as a bacterial GWAS [58]. Moreover, it should be noted that we have not collected data on the types and amount of therapeutic antibiotics used in these patients, so differences in the resistome should be interpreted with care.

A previous study based on the R-GNOSIS ICU data found that phenotypic resistance to colistin was rare [14], despite this being one of the antibiotics administered in SDD. In concordance with this result, no *mcr* genes were predicted based on WGS data, and only four isolates with phenotypic resistance were reported. Nevertheless, phenotypic resistance was determined using the E-test method, which has limited predictive accuracy according to Galani et al. [60].

Overall, our study constitutes the first WGS-based analysis of the potential effects of SDD in the pan-genome composition of a PPMO. Despite the limitations in sample collection, our results suggest that SDD treatment has limited effects in the accessory genome, plasmidome and resistome compositions of ESBL-*E. coli*.

## References

1. Stoutenbeek CP, van Saene HK, Miranda DR, Zandstra DF. The effect of selective decontamination of the digestive tract on colonisation and infection rate in multiple trauma patients. *Intensive Care Med.* 1984;10. doi:10.1007/BF00259435
2. Van der Waaij D, Berghuis-de Vries JM, Lekkerkerk-van der Wees JE. Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. *Epidemiology & Infection.* 1971;69: 405–411.
3. Wittekamp BHJ, Oostdijk EAN, Cuthbertson BH, Brun-Buisson C, Bonten MJM. Selective decontamination of the digestive tract (SDD) in critically ill patients: a narrative review. *Intensive Care Med.* 2019;46: 343–349.
4. Bergmans DCJJ, Bergmans DCJ, Bonten MJM, Gaillard CA, Paling JC, van der GEEST S, et al. Prevention of Ventilator-associated Pneumonia by Oral Decontamination. *American Journal of Respiratory and Critical Care Medicine.* 2001. pp. 382–388. doi:10.1164/ajrccm.164.3.2005003
5. WHO Regional Office for Europe/European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2022 – 2020 data. Copenhagen: WHO Regional Office for Europe; 2022. 2022.
6. Jonge E de, de Jonge E, Schultz MJ, Spanjaard L, Bossuyt PMM, Vroom MB, et al. Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial. *The Lancet.* 2003. pp. 1011–1016. doi:10.1016/s0140-6736(03)14409-1
7. Kluge S, Markewitz A, Muhl E, Putensen C, Quintel M, Sybrecht GW. DIVI Jahrbuch 2013/2014: Fortbildung und Wissenschaft in der interdisziplinären Intensivmedizin und Notfallmedizin. MWV; 2015.
8. Jonge E de, de Jonge E. Effects of selective decontamination of digestive tract on mortality and antibiotic resistance in the intensive-care unit. *Current Opinion in Critical Care.* 2005. pp. 144–149. doi:10.1097/01.ccx.0000155352.01489.11
9. Oostdijk EAN, Kesecioglu J, Schultz MJ, Visser CE, de Jonge E, van Essen EHR, et al. Effects of decontamination of the oropharynx and intestinal tract on antibiotic resistance in ICUs: a randomized clinical trial. *JAMA.* 2014;312: 1429–1437.
10. van Hout D, Plantinga NL, Bruijning-Verhagen PC, Oostdijk EAN, de Smet AMGA, de Wit GA de, et al. Cost-effectiveness of selective digestive decontamination (SDD) versus selective oropharyngeal decontamination (SOD) in intensive care units with low levels of antimicrobial resistance: an individual patient data meta-analysis. *BMJ Open.* 2019;9: e028876.
11. Plantinga NL, Wittekamp BHJ, Brun-Buisson C, Bonten MJM, R-GNOSIS ICU study group. The effects of topical antibiotics on eradication and acquisition of third-generation cephalosporin and carbapenem-resistant Gram-negative bacteria in ICU patients; a post hoc analysis from a multicentre cluster-randomized trial. *Clin Microbiol Infect.* 2020;26: 485–491.
12. European Centre for Disease Prevention and Control. Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals. Stockholm: ECDC; 2013.
13. Houben AJM, Oostdijk EAN, van der Voort PHJ, Monen JCM, Bonten MJM, van der Bij AK, et al. Selective decontamination of the oropharynx and the digestive tract, and antimicrobial resistance: a 4 year ecological study in 38 intensive care units in the Netherlands. *Journal of Antimicrobial Chemotherapy.* 2014. pp. 797–804. doi:10.1093/jac/dkt416
14. Wittekamp BH, Plantinga NL, Cooper BS, Lopez-Contreras J, Coll P, Mancebo J, et al. Decontamination Strategies and Bloodstream Infections With Antibiotic-Resistant Microorganisms in Ventilated Patients: A Randomized Clinical Trial. *JAMA.* 2018;320: 2087–2098.
15. Daneman N, Sarwar S, Fowler RA, Cuthbertson BH, SuDDICU Canadian Study Group. Effect of selective decontamination on antimicrobial resistance in intensive care units: a systematic review and meta-analysis. *Lancet Infect Dis.* 2013;13: 328–341.

16. Buelow E, González T d. j., Fuentes S, de Steenhuijsen Piters WAA, Lahti L, Bayjanov JR, et al. Comparative gut microbiota and resistome profiling of intensive care patients receiving selective digestive tract decontamination and healthy subjects. *Microbiome*. 2017. doi:10.1186/s40168-017-0309-z
17. Buelow E, Gonzalez TB, Versluis D, Oostdijk EAN, Ogilvie LA, van Mourik MSM, et al. Effects of selective digestive decontamination (SDD) on the gut resistome. *Journal of Antimicrobial Chemotherapy*. 2014. pp. 2215–2223. doi:10.1093/jac/dku092
18. Benus RF, Harmsen HJ, Welling GW, Spanjersberg R, Zijlstra JG, Degener JE, et al. Impact of digestive and oropharyngeal decontamination on the intestinal microbiota in ICU patients. *Intensive Care Med*. 2010;36: 1394–1402.
19. van Doorn-Schepens MLM, Abis GSA, Oosterling SJ, van Egmond M, Poort L, Stockmann HBA, et al. The effect of selective decontamination on the intestinal microbiota as measured with IS-pro: a taxonomic classification tool applicable for direct evaluation of intestinal microbiota in clinical routine. *Eur J Clin Microbiol Infect Dis*. 2022;41: 1337.
20. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*. 2017. p. e1005595. doi:10.1371/journal.pcbi.1005595
21. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom*. 2021;7. doi:10.1099/mgen.0.000685
22. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*. 2020;21: 180.
23. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics*. 2015;16: 1–8.
24. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res*. 2019;29: 304.
25. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics*. 2020;36: 3874–3876.
26. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*. 2016. doi:10.1186/s13059-016-0997-x
27. Arredondo-Alonso S, Gladstone RA, Pöntinen AK, Gama JA, Schürch AC, Lanza VF, et al. Consistent typing of plasmids with the mge-cluster pipeline. *bioRxiv*. 2022. p. 2022.12.16.520696. doi:10.1101/2022.12.16.520696
28. Griffith DM, Veech JA, Marsh CJ. cooccur: Probabilistic Species Co-Occurrence Analysis in R. *J Stat Softw*. 2016;69: 1–17.
29. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945. p. 80. doi:10.2307/3001968
30. Fisher RA. On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *J R Stat Soc*. 1922;85: 87.
31. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11. doi:10.1016/j.mib.2008.09.006
32. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci Adv*. 2021;7. doi:10.1126/sciadv.abe3868
33. Arredondo-Alonso S, Top J, McNally A, Puranen S, Pesonen M, Pensar J, et al. Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*. *MBio*. 2020;11. doi:10.1128/mBio.03284-19



34. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat Rev Microbiol.* 2021;19: 347–359.
35. Chen F, Lv T, Xiao Y, Chen A, Xiao Y, Chen Y. Clinical Characteristics of Patients and Whole Genome Sequencing-Based Surveillance of *Escherichia coli* Community-Onset Bloodstream Infections at a Non-tertiary Hospital in CHINA. *Front Microbiol.* 2021;12. doi:10.3389/fmicb.2021.748471
36. Paramita RI, Nelwan EJ, Fadilah F, Renesteen E, Puspandari N, Erlina L. Genome-based characterization of *Escherichia coli* causing bloodstream infection through next-generation sequencing. *PLoS One.* 2020;15: e0244358.
37. Irengé LM, Ambroise J, Bearzatto B, Durant J-F, Chirimwami RB, Gala J-L. Whole-genome sequences of multidrug-resistant *Escherichia coli* in South-Kivu Province, Democratic Republic of Congo: characterization of phylogenomic changes, virulence and resistance genes. *BMC Infect Dis.* 2019;19: 1–10.
38. Thuy TTD, Lu H-F, Kuo P-Y, Lin W-H, Lin T-P, Lee Y-T, et al. Whole-genome-sequence-based characterization of an NDM-5-producing uropathogenic *Escherichia coli* EC1390. *BMC Microbiol.* 2022;22: 1–12.
39. Forde BM, Bergh H, Cuddihy T, Hajkovicz K, Hurst T, Playford EG, et al. Clinical Implementation of Routine Whole-genome Sequencing for Hospital Infection Control of Multi-drug Resistant Pathogens. *Clin Infect Dis.* 2022;76: e1277–e1284.
40. Conway T, Cohen PS. Commensal and Pathogenic *Escherichia coli* Metabolism in the Gut. *Microbiol Spectr.* 2015;3. doi:10.1128/microbiolspec.MBP-0006-2014
41. Gagarinova A, Hosseinnia A, Rahmatbakhsh M, Istace Z, Phanse S, Moutaoufik MT, et al. Auxotrophic and prototrophic conditional genetic networks reveal the rewiring of transcription factors in *Escherichia coli*. *Nat Commun.* 2022;13: 1–16.
42. Babina AM, Surkov S, Ye W, Jerlström-Hultqvist J, Larsson M, Holmqvist E, et al. Rescue of *Escherichia coli* auxotrophy by de novo small proteins. 2023 [cited 21 Mar 2023]. doi:10.7554/eLife.78299
43. Tonkin-Hill G, Gladstone RA, Pöntinen AK, Arredondo-Alonso S, Bentley SD, Corander J. Robust analysis of prokaryotic pangenome gene gain and loss rates with Panstripe. *Genome Res.* 2023;33. doi:10.1101/gr.277340.122
44. Salamzade R, Manson AL, Walker BJ, Brennan-Krohn T, Worby CJ, Ma P, et al. Inter-species geographic signatures for tracing horizontal gene transfer and long-term persistence of carbapenem resistance. *Genome Med.* 2022;14: 1–22.
45. Klassert TE, Leistner R, Zubiria-Barrera C, Stock M, López M, Neubert R, et al. Bacterial colonization dynamics and antibiotic resistance gene dissemination in the hospital environment after first patient occupancy: a longitudinal metagenetic study. *Microbiome.* 2021;9: 1–17.
46. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. 2020 [cited 21 Feb 2023]. doi:10.7554/eLife.53886
47. Brodrick HJ, Raven KE, Kallonen T, Jamrozny D, Blane B, Brown NM, et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Med.* 2017;9: 1–11.
48. Hawkey J, Wyres KL, Judd LM, Harshegyi T, Blakeway L, Wick RR, et al. ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Med.* 2022;14. doi:10.1186/s13073-022-01103-0
49. Hu Y, Zhang H, Wei L, Feng Y, Wen H, Li J, et al. Competitive Transmission of Carbapenem-Resistant *Klebsiella pneumoniae* in a Newly Opened Intensive Care Unit. *mSystems.* 2022;7. doi:10.1128/msystems.00799-22
50. Cao G, Zhao S, Kuang D, Hsu C-H, Yin L, Luo Y, et al. Geography shapes the genomics and antimicrobial resistance of *Salmonella enterica* Serovar Enteritidis isolated from humans. *Sci Rep.* 2023;13: 1–13.

51. Yu L, Wang D, Li P, Cai Y, Zhang X, Luo X, et al. Epidemiology, molecular characterization, and drug resistance of IncHI5 plasmids from Enterobacteriaceae. *Int Microbiol.* 2022; 1–8.
52. Mbelle NM, Feldman C, Sekyere JO, Maningi NE, Modipane L, Essack SY. The Resistome, Mobilome, Virulome and Phylogenomics of Multidrug-Resistant *Escherichia coli* Clinical Isolates from Pretoria, South Africa. *Sci Rep.* 2019;9. doi:10.1038/s41598-019-52859-2
53. Livermore DM, Day M, Cleary P, Hopkins KL, Toleman MA, Wareham DW, et al. OXA-1  $\beta$ -lactamase and non-susceptibility to penicillin/ $\beta$ -lactamase inhibitor combinations among ESBL-producing *Escherichia coli*. *J Antimicrob Chemother.* 2019;74: 326–333.
54. Varani A, He S, Siguier P, Ross K, Chandler M. The IS6 family, a clinically important group of insertion sequences including IS26. *Mob DNA.* 2021;12: 1–18.
55. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews.* 2018. doi:10.1128/cmr.00088-17
56. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, 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. doi:10.1073/pnas.2008731118
57. Harmer CJ, Hall RM. IS26-Mediated Formation of Transposons Carrying Antibiotic Resistance Genes. *mSphere.* 2016;1. doi:10.1128/mSphere.00038-16
58. Harmer CJ, Moran RA, Hall RM. Movement of IS26-associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS26 and preferentially inserts adjacent to another IS26. *MBio.* 2014;5. doi:10.1128/mBio.01801-14
59. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol.* 2020;10. doi:10.3389/fmicb.2019.03119
60. Galani I, Kontopidou F, Souli M, Rekatsina PD, Koratzanis E, Deliolanis J, et al. Colistin susceptibility testing by Etest and disk diffusion methods. *Int J Antimicrob Agents.* 2008;31. doi:10.1016/j.ijantimicag.2008.01.011

## Supplementary Materials

### Supplementary Data

Supplementary data can be downloaded from: <https://doi.org/10.5281/zenodo.7926668>

### Supplementary Tables

**Supplementary Table S1.** Results of PERMANOVA analysis to model the variance observed in accessory genome composition. The “\*” symbol indicates interaction term between variables. Treatment codes for SDD vs baseline.

Model nr.	Explanatory variables	Df	F.Model	R2	Pr(>F)
1	Phylogroup	7	13.5281039	0.4390271864	0.001
2	Hospital	4	1.216606935	0.03776334787	0.154
3	Treatment	1	1.365392476	0.01063676471	0.149
4	Phylogroup	7	13.7654975	0.4390271864	0.001
	Treatment	1	1.484859951	0.00676529839	0.102
	Phylogroup*Treatment	5	1.327693956	0.03024610429	0.054
5	Phylogroup	7	14.31769118	0.4390271864	0.001
	Hospital	4	1.430689111	0.02506834405	0.054
	Phylogroup*Hospital	18	1.296643498	0.102238266	0.006
6	Treatment	1	1.372662651	0.01063676471	0.143
	Hospital	4	1.2108664	0.03753202137	0.148
	Treatment*Hospital	4	0.9581897992	0.0297000561	0.532

**Supplementary Table S2.** Results of PERMANOVA analysis to model the variance observed in plasmidome composition. The “\*” symbol indicates interaction term between variables. Treatment codes for SDD vs baseline.

Model nr.	explanatory_variable	Df	F.Model	R2	Pr(>F)
1	Phylogroup	7	2.262016978	0.115717622	0.001
2	Hospital	4	1.182226021	0.03673537126	0.099
3	Treatment	1	1.413281789	0.01100572908	0.072
4	Phylogroup	7	2.311171098	0.115717622	0.001
	Treatment	1	1.346775064	0.009633063636	0.083
	Phylogroup*Treatment	5	1.456516246	0.05209004107	0.003
5	Phylogroup	7	2.362907434	0.115717622	0.001
	Hospital	4	1.346915624	0.03769251651	0.016
	Phylogroup*Hospital	18	1.222732122	0.1539779659	0.002
6	Treatment	4	1.18561335	0.01183960017	0.085
	Hospital	1	1.528465622	0.03673537126	0.042
	Treatment*Hospital	4	0.9567051965	0.02964281785	0.592

## Chapter 5

Supplementary Table S3. Fisher's test results for abundance of different COG types across study periods.

COG	p.value	Confidence interval	OR	Occurrence in SDD	Occurrence in baseline period	Not occurrence SDD	Not occurrence baseline period
J	0.9852496345	c(0.975789492902574, 1.02425758983737)	0.9997054997	13722	14429	200445	210716
K	0.6056192133	c(0.972489422695379, 1.0163625709264)	0.9941842403	16759	17713	197408	207432
F	0.9719039845	c(0.965246095111443, 1.03455419087637)	0.9993147717	6486	6823	207681	218322
Q	0.9366409714	c(0.951155658340286, 1.05601402039197)	1.002227294	2819	2957	211348	222188
G	0.6018505376	c(0.985794356531915, 1.02508083454845)	1.005237299	22074	23097	192093	202048
I	0.6933838823	c(0.974037533976659, 1.0405276915715)	1.006727248	7196	7516	206971	217629
E	0.8130938607	c(0.982364983479942, 1.02289454010526)	1.00244109	20432	21432	193735	203713
C	0.9634520666	c(0.978245285628295, 1.02330688097101)	1.000524396	16079	16895	198088	208250
M	0.8085665345	c(0.974179799777156, 1.02052782175703)	0.9970788874	15003	15815	199164	209330
O	0.906893496	c(0.969951310487423, 1.02736934293829)	0.9982685969	9547	10053	204620	215092
R	0.8888010572	c(0.972289137890958, 1.02457843874712)	0.9980901027	11625	12243	202542	212902
U	0.8477019785	c(0.952154529987101, 1.04089787909678)	0.99555203	3884	4101	210283	221044
T	0.9945601485	c(0.973350466549645, 1.02703215378781)	0.9998562239	11065	11634	203102	213511
P	0.9345003041	c(0.976456561365551, 1.026244226516)	1.001042395	13000	13653	201167	211492
H	0.8402077964	c(0.976601630259693, 1.02963648049424)	1.002776834	11422	11976	202745	213169
S	0.9486169264	c(0.96991901626744, 1.03340134536923)	1.001155743	7806	8197	206361	216948
D	0.8987001219	c(0.949242121483854, 1.06140674472677)	1.003776748	2471	2588	211696	222557
V	0.9926284441	c(0.963969778713874, 1.03672824250224)	0.9996686522	5875	6178	208292	218967
L	0.5763402559	c(0.960966481318246, 1.02237950636188)	0.9912170141	8154	8645	206013	216500
X	0.0649580104	c(0.996571842983943, 1.11875968712521)	1.055920887	2362	2353	211805	222792

Impact of selective digestive decontamination on the pangenome composition of ESBL-*E. coli*

Supplementary Table S3. Cont.

COG	p.value	Confidence interval	OR	Occurrence in SDD	Occurrence in baseline period	Not occurrence SDD	Not occurrence baseline period
N	0.1746329295	c(0.934173122750306, 1.01245725611287)	0.9725566337	4723	5102	209444	220043
W	0.970997954	c(0.928852378768879, 1.07256736623477)	0.9981455043	1485	1564	212682	223581
A	0.6192491546	c(0.740511742054748, 1.68314621513269)	1.115649109	52	49	214115	225096
Z	1	c(0.779773374511084, 1.29089404928233)	1.003476705	126	132	214041	225013

## Chapter 5

**Supplementary Table S4.** Co-occurrence of ARGs in the same plasmid in SDD isolates. Co-occurrences with a p-value smaller than 0.01 were considered significant.

ARG 1	ARG 2	ARG 1 Occurrence	ARG 2 Occurrence	Co-occurrence	Prob. co-occurrence	Expected co-occurrence	p.value
aac(3)-Ild	blaTEM-1	7	21	4	0.014	1.4	0.03039
aac(3)-Ild	mph(A)	7	24	4	0.016	1.6	0.04957
aac(3)-Ile	aac(6')-Ib-cr5	11	13	4	0.013	1.4	0.0312
aac(3)-Ile	blaCTX-M-15	11	18	8	0.019	1.9	2.00E-05
aac(3)-Ile	blaOXA-1	11	14	4	0.015	1.5	0.04108
aac(3)-Ile	catB3	11	14	4	0.015	1.5	0.04108
aac(6')-Ib-cr5	blaCTX-M-15	13	18	10	0.022	2.3	0
aac(6')-Ib-cr5	blaOXA-1	13	14	13	0.017	1.8	0
aac(6')-Ib-cr5	catB3	13	14	13	0.017	1.8	0
aac(6')-Ib-cr5	mph(A)	13	24	7	0.029	3	0.01073
aadA1	aadA2	17	8	5	0.013	1.3	0.00286
aadA1	aadA5	17	25	0	0.04	4.1	1
aadA1	blaCTX-M-15	17	18	0	0.029	3	1
aadA1	dfrA17	17	31	0	0.05	5.1	1
aadA1	mph(A)	17	24	0	0.038	4	1
aadA5	blaCTX-M-15	25	18	9	0.042	4.4	0.00842
aadA5	blaOXA-1	25	14	7	0.033	3.4	0.02319
aadA5	blaTEM-1	25	21	9	0.049	5.1	0.02964
aadA5	catB3	25	14	7	0.033	3.4	0.02319
aadA5	dfrA17	25	31	20	0.073	7.5	0
aadA5	mph(A)	25	24	17	0.057	5.8	0
aadA5	sul1	25	33	21	0.078	8	0
aph(3'')-Ib	aph(6)-Ild	19	19	19	0.034	3.5	0
aph(3'')-Ib	blaTEM-1	19	21	10	0.038	3.9	0.00048
aph(3'')-Ib	sul2	19	23	17	0.041	4.2	0
aph(3'')-Ib	tet(A)	19	20	10	0.036	3.7	0.00028
aph(6)-Ild	blaTEM-1	10	21	10	0.038	3.9	0.00048
aph(6)-Ild	sul2	19	23	17	0.041	4.2	0
aph(6)-Ild	tet(A)	19	20	10	0.036	3.7	0.00028
blaCTX-M-15	blaOXA-1	18	14	11	0.024	2.4	0
blaCTX-M-15	catB3	18	14	11	0.024	2.4	0
blaCTX-M-15	dfrA17	18	31	9	0.053	5.4	0.04365
blaCTX-M-15	sul1	18	33	10	0.056	5.8	0.02113
blaOXA-1	catB3	14	14	14	0.018	1.9	0
blaOXA-1	mph(A)	14	24	7	0.032	3.3	0.01802
blaOXA-1	sul1	14	33	8	0.044	4.5	0.03464
blaTEM-1	sul1	21	33	11	0.065	6.7	0.02609
catB3	mph(A)	14	24	7	0.032	3.3	0.01802
catB3	sul1	14	33	8	0.044	4.5	0.03464

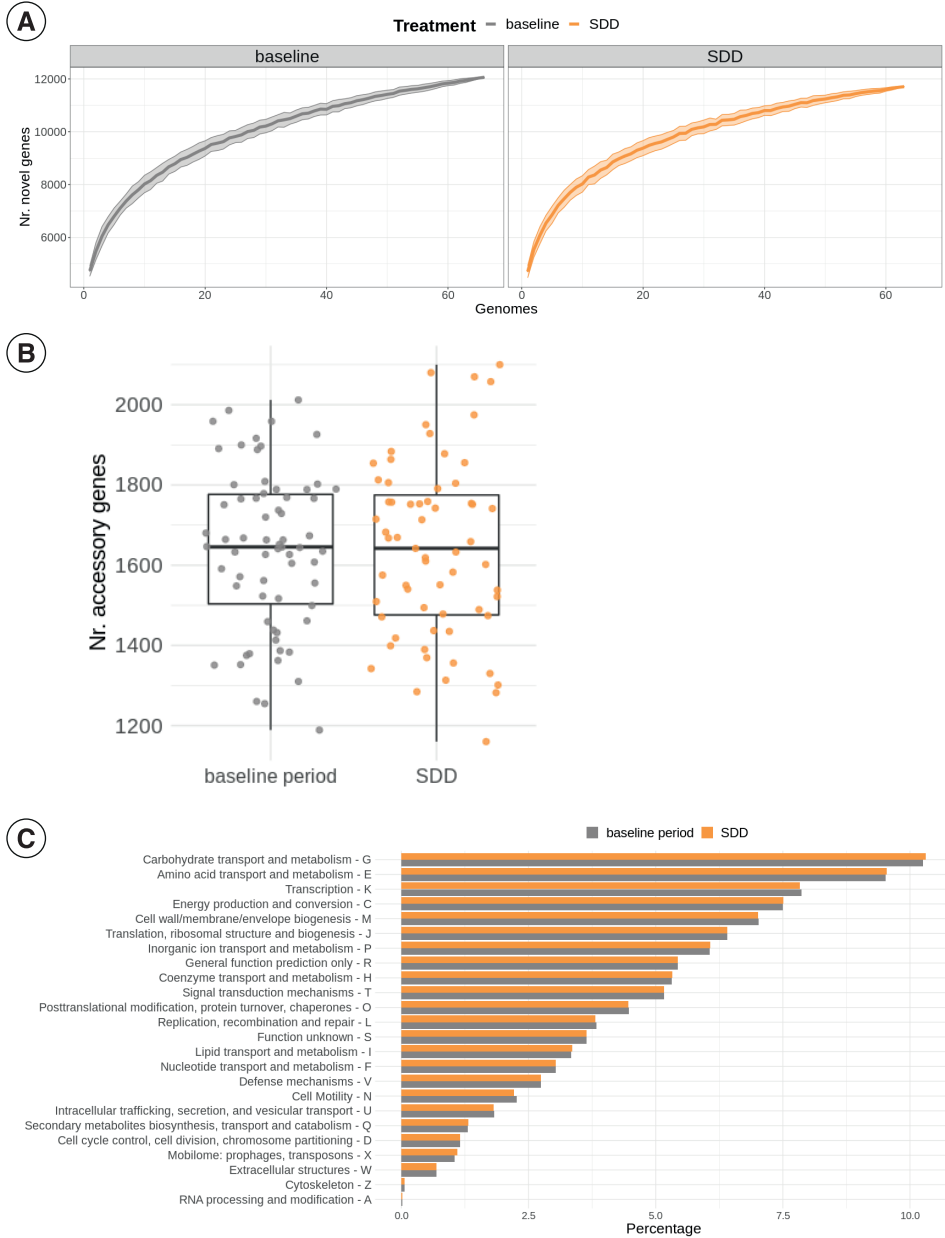
Supplementary Table S4. Cont.

ARG 1	ARG 2	ARG 1 Occurrence	ARG 2 Occurrence	Co-occurrence	Prob. co-occurrence	Expected co-occurrence	p.value
dfrA17	mph(A)	31	24	12	0.07	7.2	0.01661
dfrA17	sul1	31	33	16	0.096	9.9	0.00573
mph(A)	sul1	24	33	18	0.075	7.7	0
sul2	tet(A)	23	20	9	0.043	4.5	0.01047

Supplementary Table S5. Co-occurrence of ARGs in the same plasmid in baseline isolates. Co-occurrences with a p-value smaller than 0.01 were considered significant.

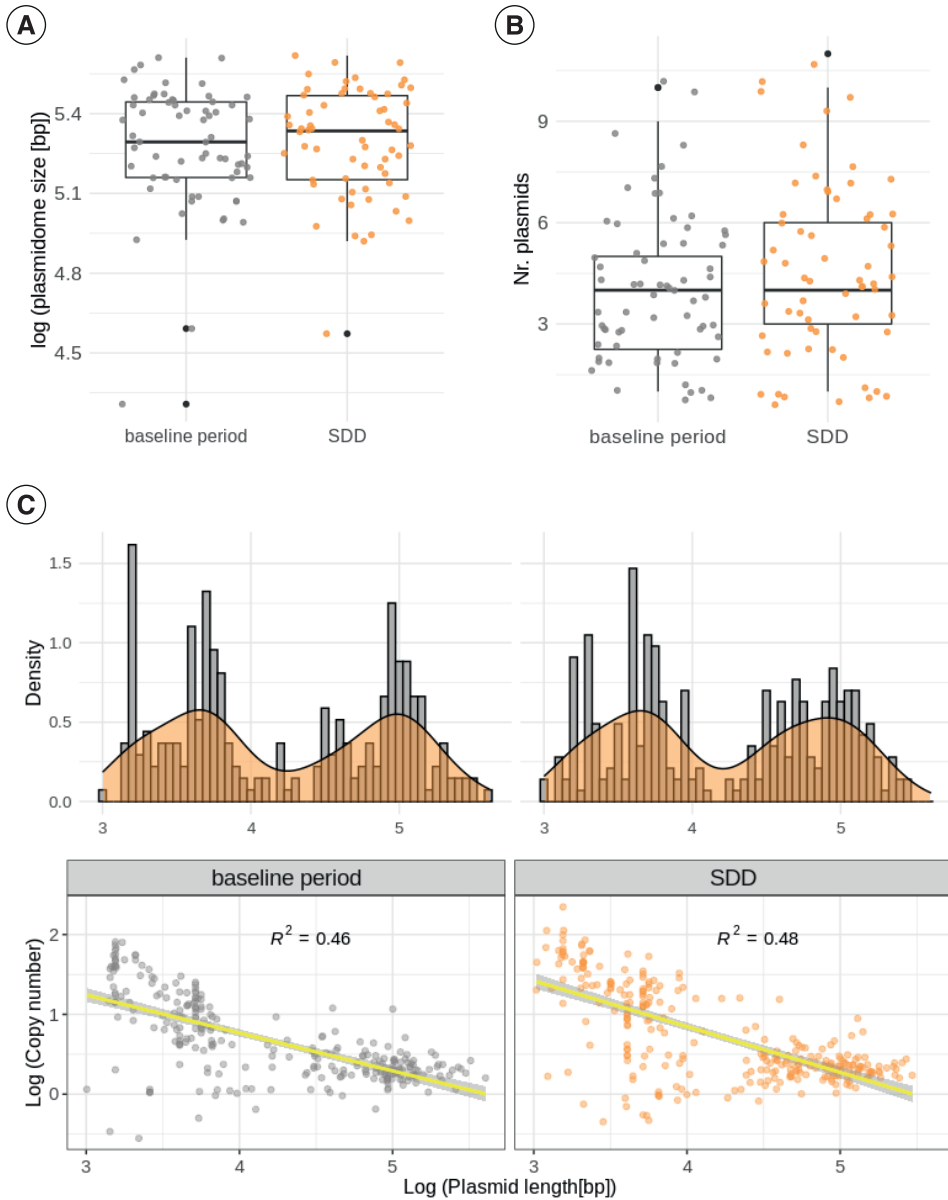
ARG 1	ARG 2	ARG 1 Occurrence	ARG 2 Occurrence	Co-occurrence	Prob. co-occurrence	Expected co-occurrence	p.value
aadA1	aadA2	14	10	4	0.011	1.2	0.02122
aadA1	aadA5	14	22	0	0.025	2.8	1
aadA1	dfrA17	14	29	0	0.032	3.6	1
aadA2	dfrA17	10	29	0	0.023	2.6	1
aadA5	dfrA17	22	29	11	0.051	5.7	0.00601
aadA5	mph(A)	22	20	12	0.035	3.9	1.00E-05
aadA5	sul1	22	31	14	0.054	6.1	8.00E-05
aadA5	sul2	22	38	12	0.067	7.5	0.02302
aph(3'')-lb	aph(3')-la	39	12	9	0.037	4.2	0.00339
aph(3'')-lb	aph(6)-ld	39	39	39	0.121	13.6	0
aph(3'')-lb	blaTEM-1	39	41	20	0.127	14.3	0.01618
aph(3'')-lb	dfrA1	39	4	4	0.012	1.4	0.01324
aph(3'')-lb	dfrA17	39	29	3	0.09	10.1	0.99989
aph(3'')-lb	sul2	39	38	32	0.118	13.2	0
aph(3')-la	aph(6)-ld	12	39	9	0.037	4.2	0.00339
aph(3')-la	blaTEM-1	12	41	9	0.039	4.4	0.00522
aph(3')-la	sul2	12	38	9	0.036	4.1	0.0027
aph(6)-ld	blaTEM-1	39	41	20	0.127	14.3	0.01618
aph(6)-ld	dfrA1	39	4	4	0.012	1.4	0.01324
aph(6)-ld	dfrA17	39	29	3	0.09	10.1	0.99989
aph(6)-ld	sul2	39	38	32	0.118	13.2	0
blaCTX-M-27	sul1	4	31	4	0.01	1.1	0.00507
blaTEM-1	dfrA17	41	29	5	0.095	10.6	0.99784
blaTEM-1	sul2	41	38	21	0.124	13.9	0.00333
blaTEM-1	sul3	41	6	5	0.02	2.2	0.02412
blaTEM-1	tet(M)	41	3	3	0.01	1.1	0.04677
dfrA1	sul2	4	38	4	0.012	1.4	0.01188
dfrA17	tet(A)	29	35	4	0.081	9.1	0.9968
mph(A)	sul1	20	31	15	0.049	5.5	0
mph(A)	tet(A)	20	35	10	0.056	6.2	0.04451
tet(A)	tet(B)	35	13	1	0.036	4.1	0.99459

Supplementary Figures

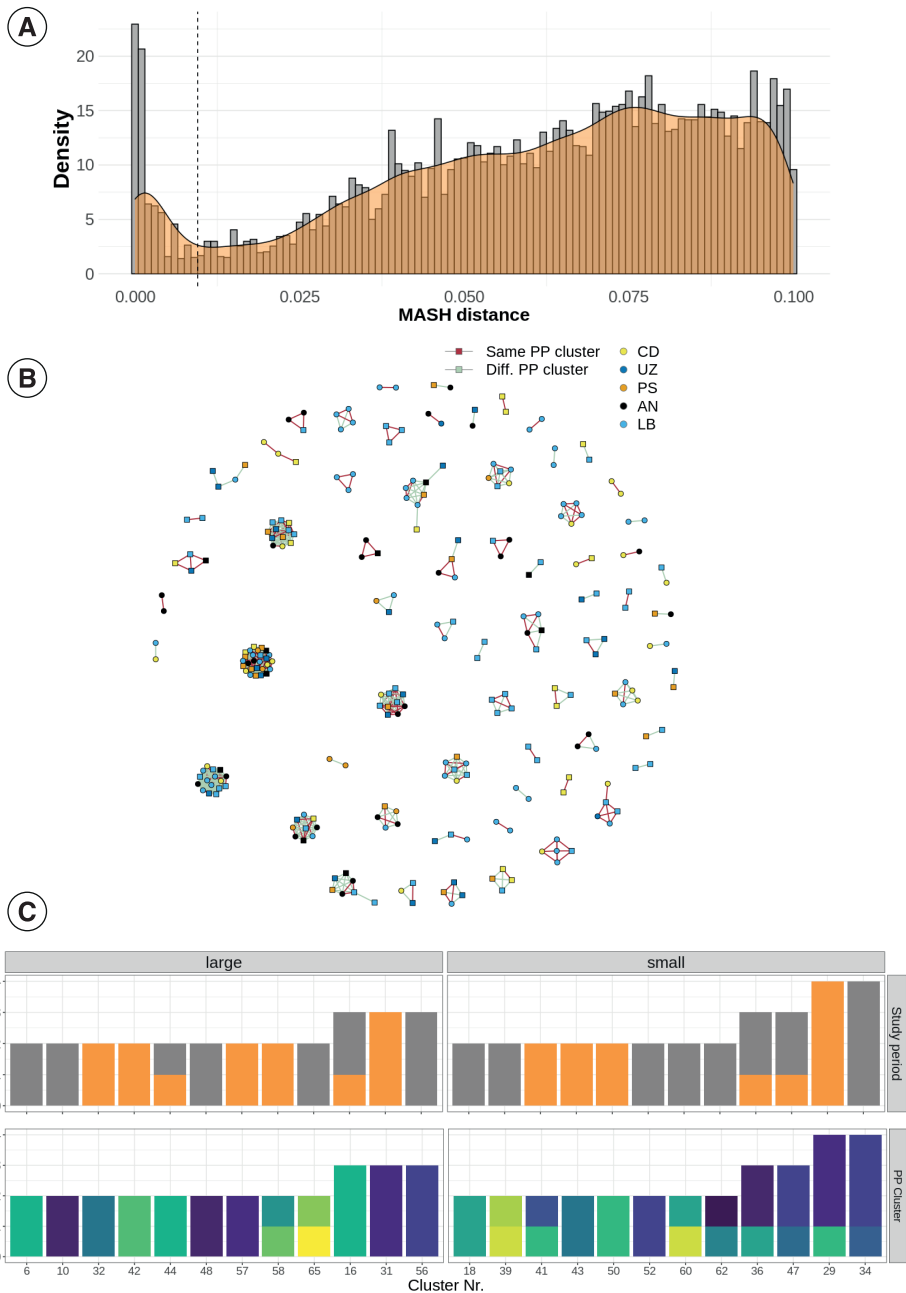


Supplementary Figure S1. A) Pangenome accumulation curves for baseline and SDD isolates. B) Number of accessory genes per isolate across different study periods. C) Fraction of COG functional categories by study period.

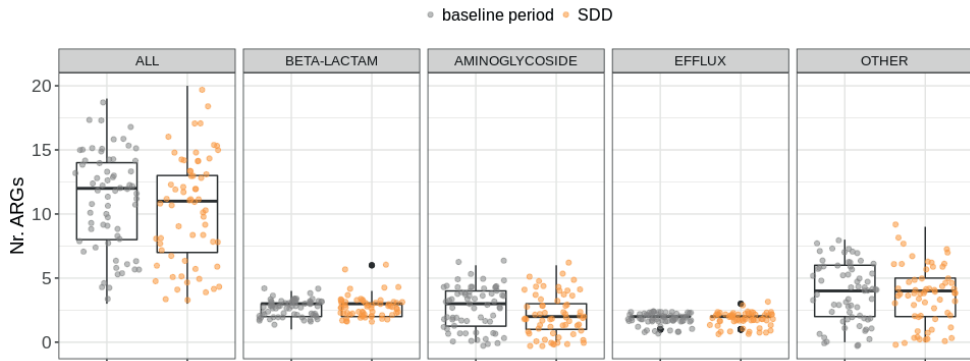




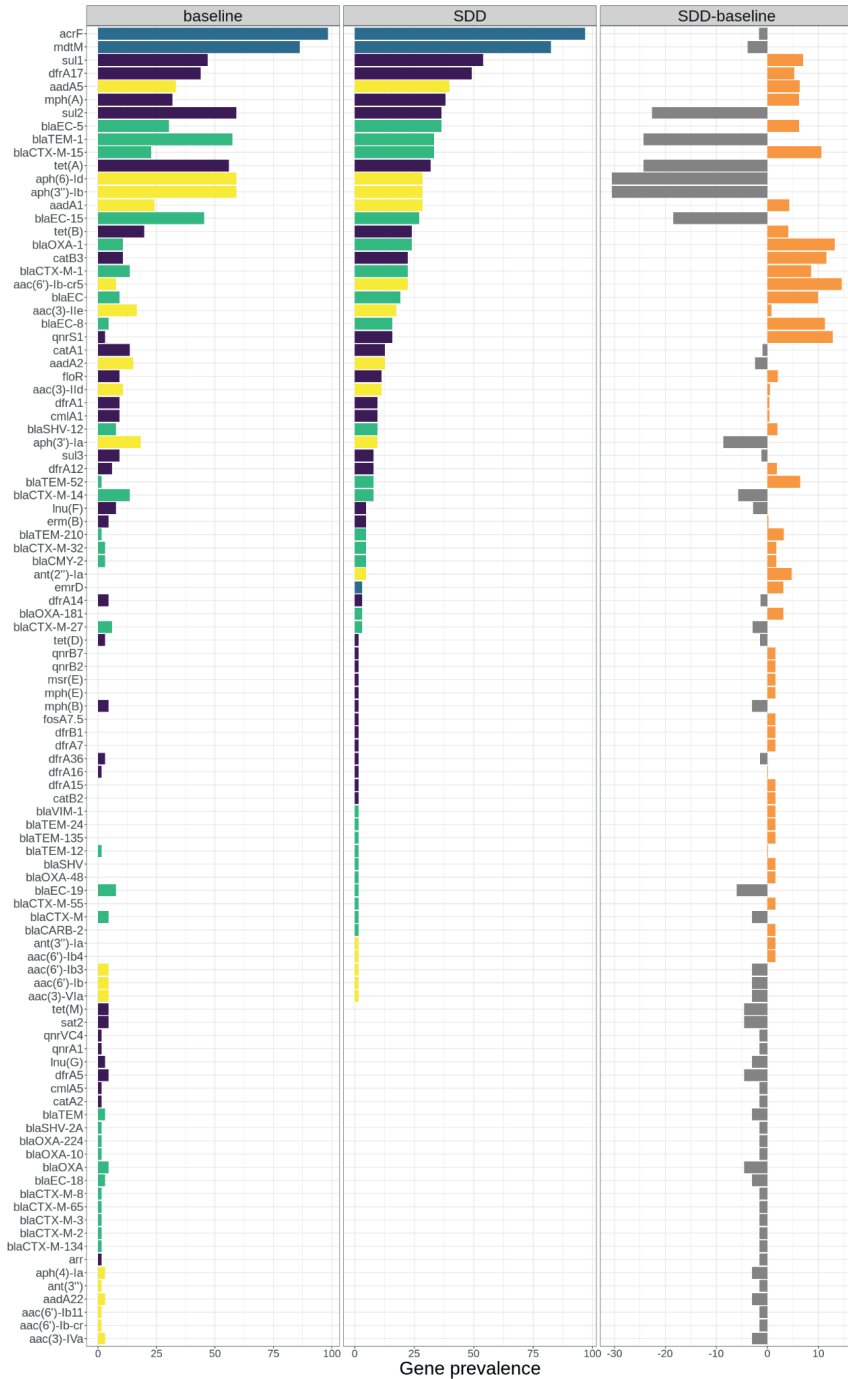
**Supplementary Figure S2.** A) Total size of predicted plasmidome sorted by study period. This is, the total length of all contigs within an isolate predicted to be plasmid by plasmidEC. B) Number of predicted individual plasmids per isolate. Plasmids were predicted using gplas. C) Plasmid size vs estimated copy number for individual plasmid predictions. In the top panel, the y-axis shows Kernell probability density function values, based on the abundance of the different plasmid sizes.



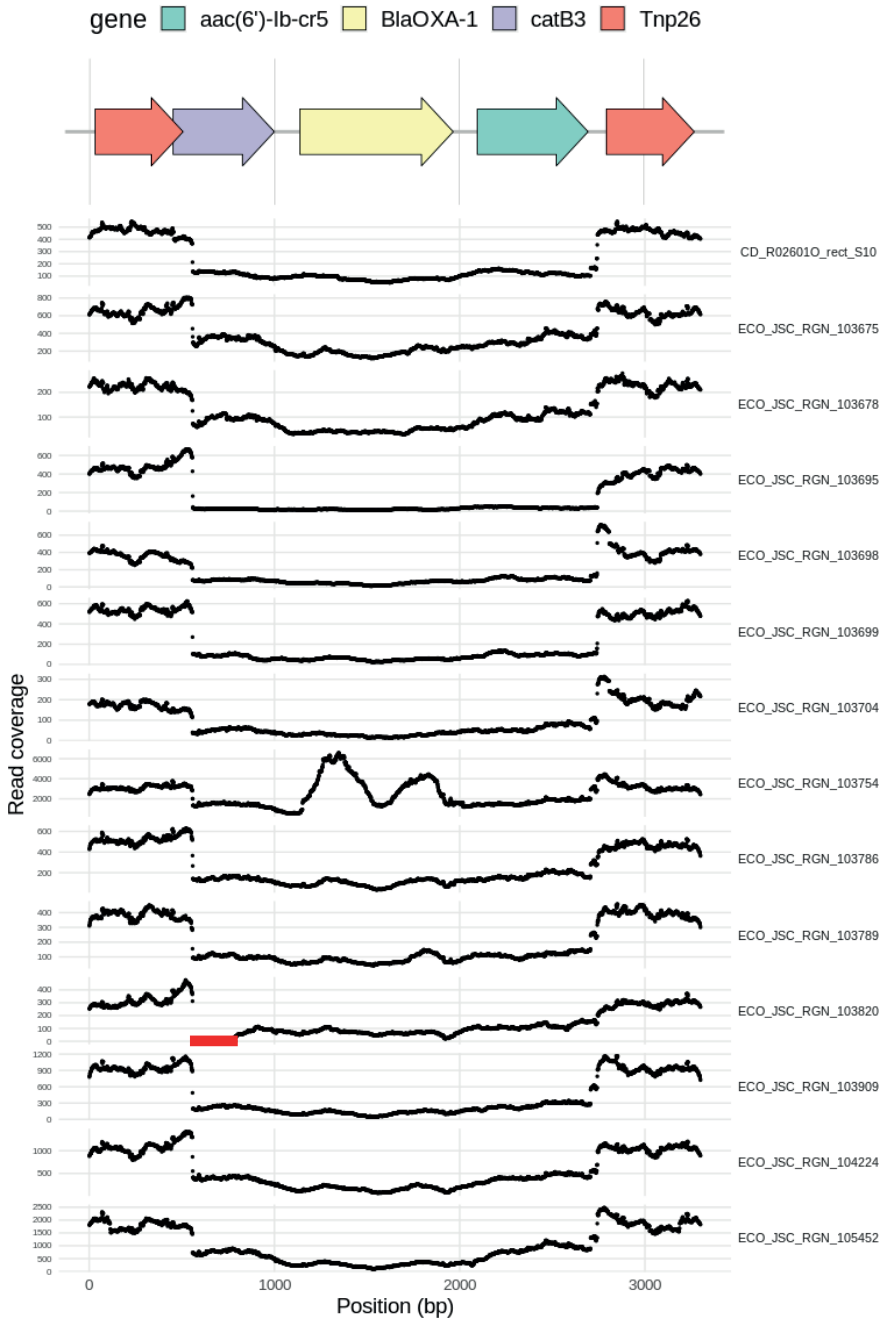
**Supplementary Figure S3.** A) MASH distances ( $k=21, s=10,000$ ) for all plasmid-predictions vs plasmid predictions. Dashed line indicates the cut-off point (distance = 0.01) to create network of plasmids B) Network displaying clusters of highly similar plasmids. C) Histograms of plasmid clusters of large plasmids and small plasmids that were present only in single hospitals, colour coded by study period (above) and by PopPUNK (PP) clusters (below).



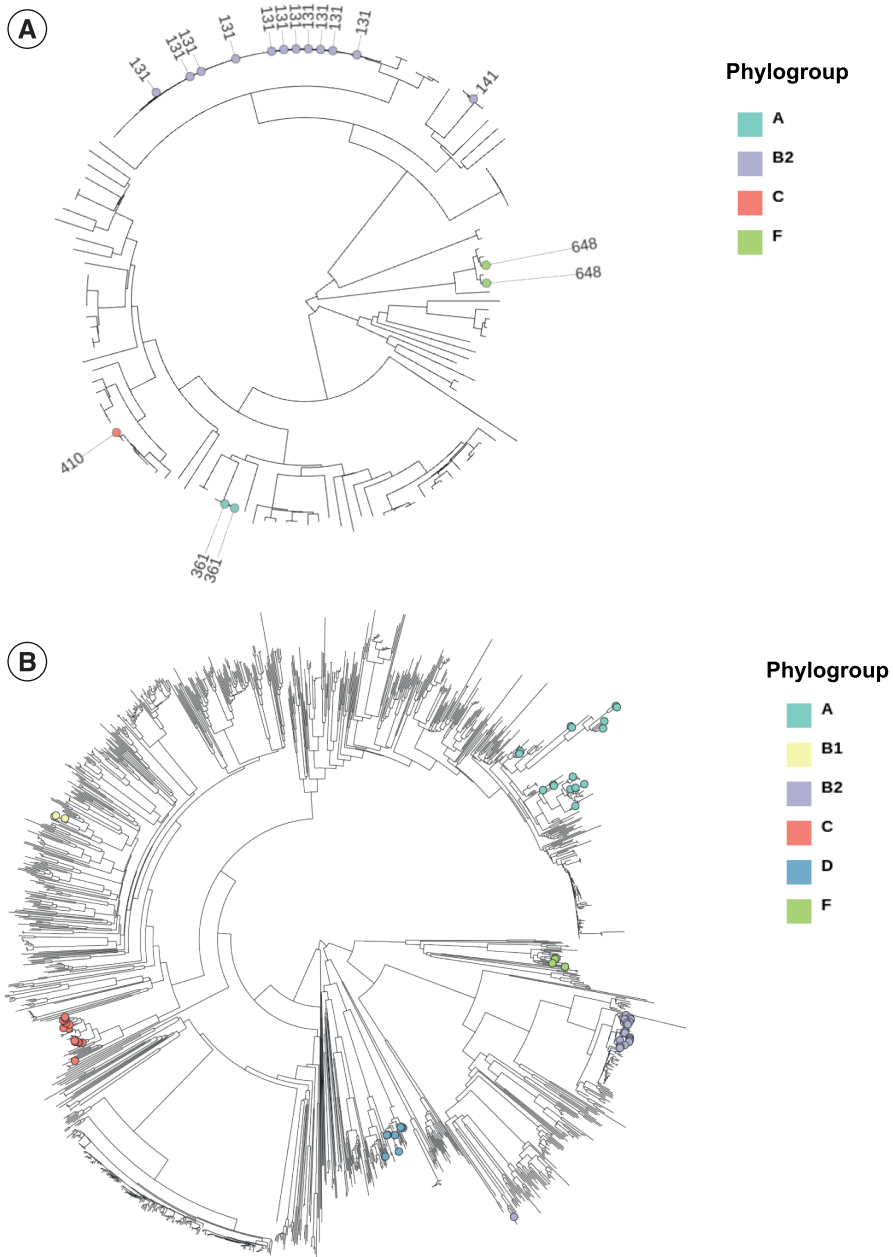
Supplementary Figure S4. Nr. of acquired ARGs per isolate, treatment and ARG type.



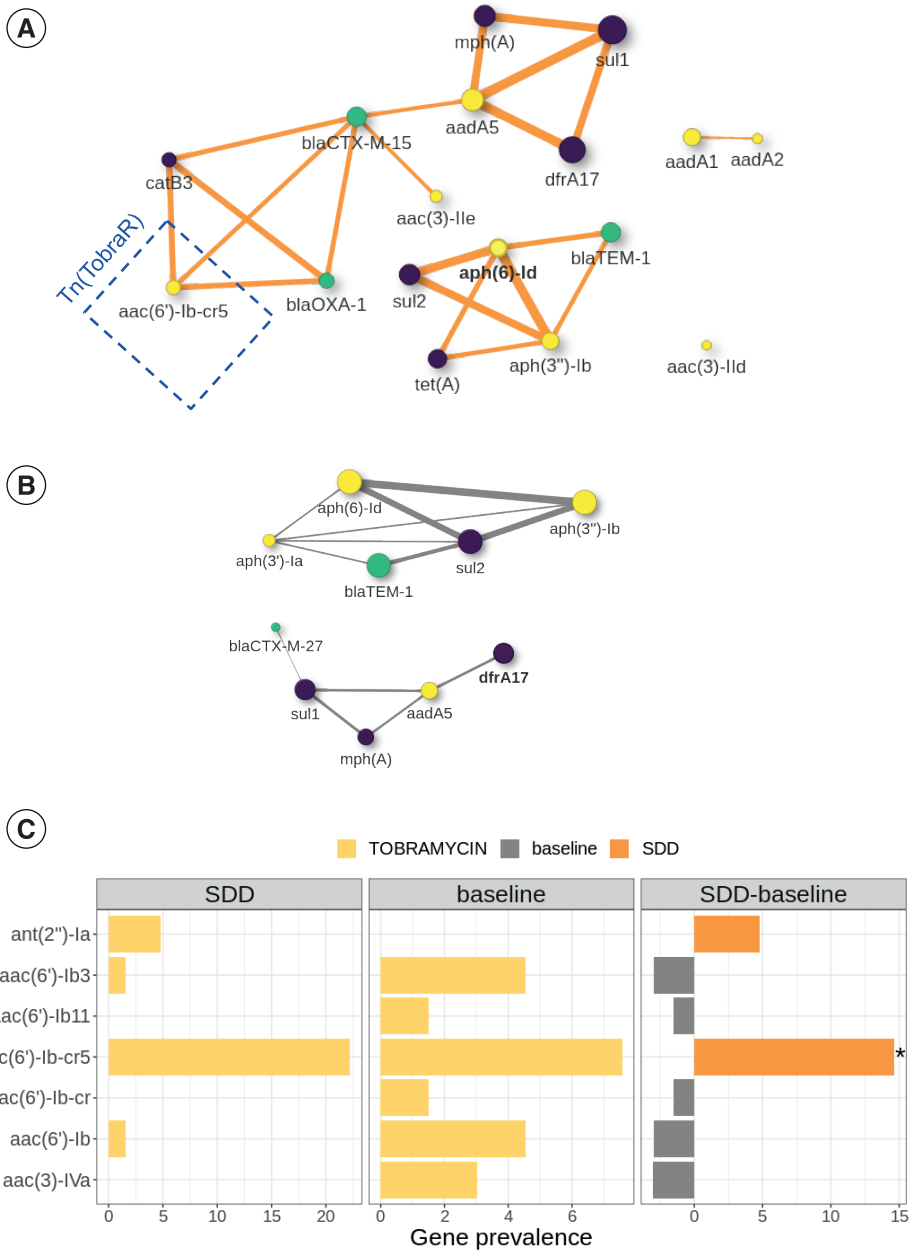
Supplementary Figure S5. The first two panels show the prevalence of all acquired ARGs in SDD and baseline isolates. The third panel shows the absolute difference between these prevalences in SDD and baseline periods.



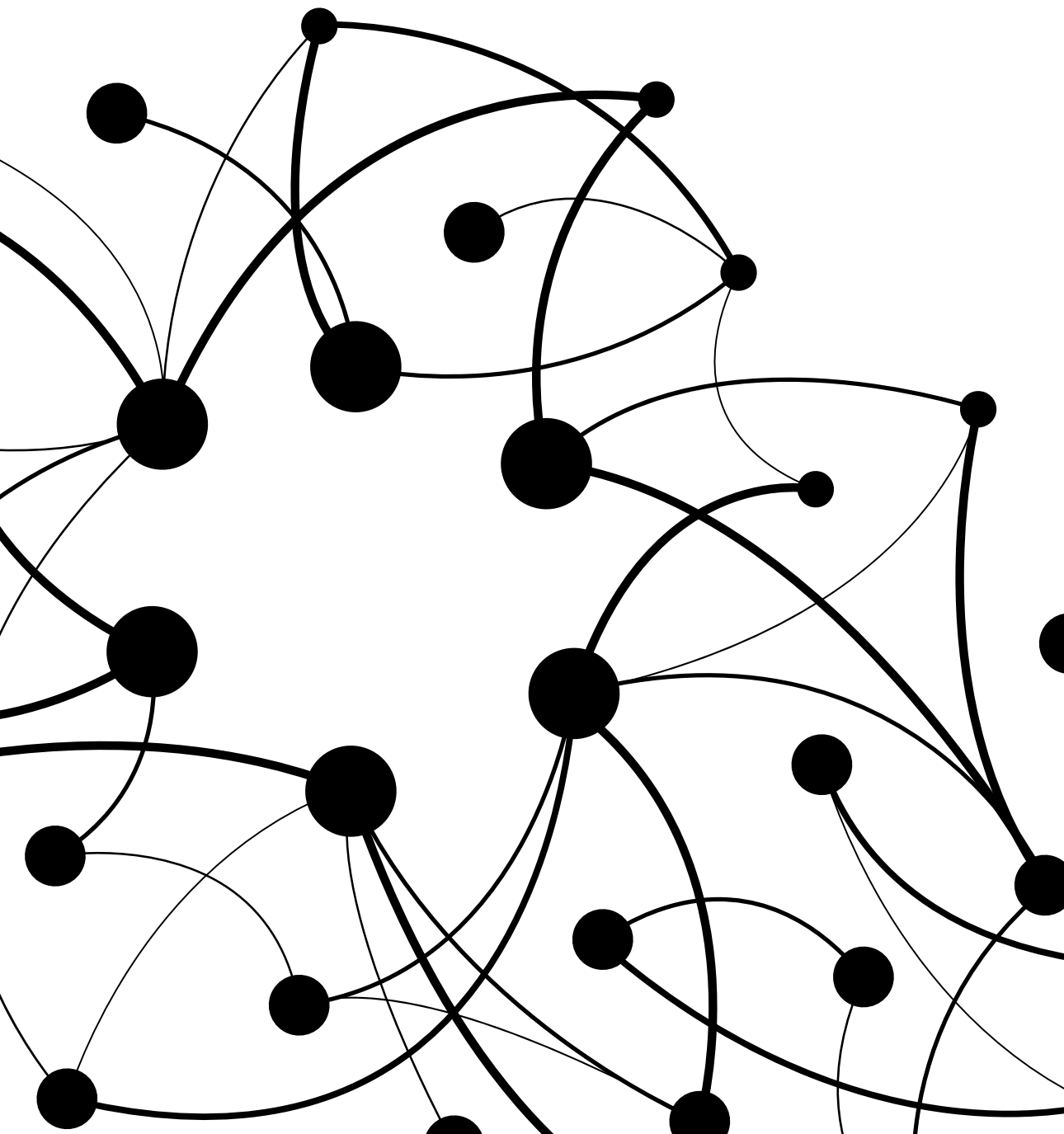
Supplementary Figure S6. Read coverage of each SDD isolate ( $n=14$ ) that contains the putative transposon carrying the tobramycin resistance gene. Coverage suggests that all SDD isolates, except ECO-JSC-RGN-103820, carry the complete sequence of Tn(TobraR). Red lines indicate regions with read coverage equal to zero.



**Supplementary Figure S7.** **A)** NJ tree based on core-genome alignment of *E. coli* isolates from the R-GNOSIS study. Leaf labels indicate the sequence type of isolates. **B)** NJ cg-tree based on k-mer presence/absence of 1381 publicly available *E. coli* complete genomes. In both trees, colored nodes indicate the genomes that carry Tn(TobraR) and their corresponding phylogroup.



**Supplementary Figure S8.** A) Co-occurrence network of ARGs in the same plasmid prediction in SDD isolates. B) Co-occurrence network of ARGs in the same plasmid prediction in baseline isolates. Only connections with a  $p$ -value  $\leq 0.01$  are drawn. C) The first two panels show the prevalence of all acquired ARGs that are predicted to provide resistance to tobramycin in baseline and SDD isolates. The third panel shows the absolute difference between prevalences in both study periods, stars indicate genes that are significantly associated with a study period (Fisher's exact test).

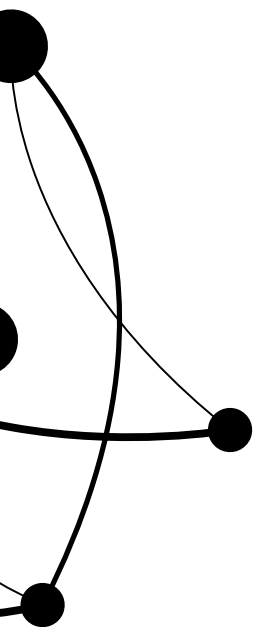




# 06

---

## Summary and General Discussion



## Summary and General Discussion

In this thesis, I aimed to assess the performance of existing plasmid prediction tools and to develop new methods to improve reconstruction of plasmids from short reads, with an emphasis on antimicrobial resistance (AMR) plasmids. These techniques were later applied to compare the pangenome composition and occurrence of AMR plasmids in ESBL-*E. coli* isolates, obtained from surveillance cultures of patients treated with selective digestive decontamination (SDD) and patients not treated with SDD in a multinational cluster-randomised study.

Below, I will present an overview of the results presented in **chapters 2-5** and discuss the advantages and limitations of the tools developed in **chapter 3** and **4**. Moreover, I will provide directions for future applications of these methods in the clinic and for AMR surveillance.

### The starting point

In **chapter 2** of this thesis we reviewed all available tools to predict plasmids from short-read data. We found a total of 25 tools and most of them ( $n=21$ ) were categorised as binary classification tools ( $n=10$ ), if the output was limited at classifying the origin of contigs as ‘plasmid’ or ‘chromosome’; or plasmid reconstruction tools ( $n=11$ ) when the tools aimed at identifying individual plasmids in a genome. Additionally, the computational (and biological) foundations behind each tool were described, along with their strongpoints and weaknesses. Since the publication of this work, in 2021, multiple other plasmid prediction tools have been developed and published [1–4], confirming the sustained interest of the scientific community in improving plasmid prediction methods using short-reads.

The previous study describing an independent comparison between plasmid prediction tools was published in 2017 [5], and it did not include most of the newly developed tools. Consequently, we decided to **benchmark the performance of six plasmid reconstruction tools**, when applied to a dataset of 240 *Escherichia coli* genomes obtained from public databases. The decision of focusing this benchmark on *E. coli* was motivated by its clinical relevance as a resistant pathogen [6,7], by the essential role that plasmids have in the dissemination of AMR genes within this species [8,9] and by the clinical aim of this thesis, which is covered in **chapter 5**. Despite the narrow scope of this study, the benchmark dataset constitutes one of its main strengths, because the selected genomes captured most of the phylogenetic and plasmidome diversity of *E. coli* and included isolates from multiple ecological niches. Our results revealed that of the six plasmid reconstruction tools, MOB-suite correctly reconstructed the largest fraction of plasmids (50%) and also identified the majority of plasmid-borne AMR genes (89%). Based on these results, we classified MOB-suite as the best plasmid reconstruction tool, and plasmidSPAdes as second best. The publication of this study has guided other researchers into selecting MOB-suite to perform their own plasmidome analysis using short reads [10,11]. Importantly, this study also highlighted that **all tools had major difficulties at reconstructing AMR-plasmids of *E. coli***.

### Why are AMR plasmids so difficult to reconstruct using short-reads?

We hypothesised that the correct reconstruction of AMR-plasmids using short reads constituted a particularly hard challenge due to multiple reasons. First, AMR carrying contigs are usually short [12], highly conserved [13], and associated with multiple plasmid backbones [14,15]. Consequent-

ly, assigning these contigs to the correct plasmid type is problematic when tools solely rely on reference databases without considering the immediate genomic context of each contig. As support for this claim, in **chapter 3** we found that MOB-suite fragmented single AMR-plasmids into multiple plasmid types. Second, AMR-plasmids usually have a low copy number [16,17], which causes contigs to have a sequencing coverage similar to the chromosome. Therefore, tools that are based on the assumption that plasmids should have a higher coverage [18,19], struggle to correctly identify AMR-plasmid contigs. The latter was confirmed in **chapter 4** by the observation that plasmidSPAdes predictions had low recall and precision when the estimated copy number of plasmids approached one.

### Gplas & plasmidEC: Towards an improved reconstructions of AMR plasmids in *E. coli*

In **chapter 3** we aimed at improving the reconstruction of AMR-plasmids in *E. coli*. We speculated that these predictions could be improved by exploiting the contig connectivity information that is embedded in assembly graphs, as implemented in gplas [20]. This tool had proven very successful when reconstructing plasmids of *Enterococcus faecium*, but frequently generated incomplete AMR-plasmid predictions in *E. coli*, as indicated by the low recall (or completeness) values obtained in our benchmark study (median=0.29, IQR=0.14 - 0.62) of **chapter 2**. We tackled this issue by introducing two modifications to gplas. First, we developed a *bold* mode that improves the binning of plasmids with pronounced sequencing coverage variations. Second, we replaced mlplasmids [21] with plasmidEC for identification of plasmid nodes in the assembly graph. The integration of plasmidEC into gplas was essential for improving plasmid reconstructions. PlasmidEC correctly identified 94% of all contigs derived from AMR-plasmids and considerably outperformed mlplasmids, which identified 73% of these. PlasmidEC is an ensemble of three binary classification tools (for *E. coli* these are: PlaScope, Platon, RFPlasmid) that implements a majority vote system to predict the origin of contigs. The development of this tool was motivated by two observations derived from a comparison of four binary classification tools (**Chapter 3**): (1) 95% of plasmid contigs were correctly classified by at least two different tools, and (2) most misclassifications of plasmid contigs (55%) were made by a single tool.

Our new method, gplas\_plasmidEC, generated predictions that included large fractions of AMR-plasmids, as indicated by the high completeness(bp) values (median=0.82, IQR= 0.52 - 0.92), that exceeded those of MOB-suite (median=0.32, IQR=0.11 - 0.80). We found that MOB-suite fragmented 49% of AMR-plasmids into multiple predictions, while gplas\_plasmidEC did so in only 14% of the cases. Additionally, our tool had higher values of accuracy(bp), suggesting that a smaller number of chimeric predictions were generated. Finally, both tools detected the same number of plasmid-borne AMR genes. These results indicated that gplas\_plasmidEC is currently the best available method to reconstruct AMR-plasmids of *E. coli* from short-reads.

### What about other species?

Given the results obtained for *E. coli*, in **chapter 4** we expanded the range of species of plasmidEC. We built four species-specific binary classifiers, aimed at classifying contigs of clinically relevant pathogens (*Enterococcus faecium*, *Klebsiella pneumoniae*, *Salmonella enterica* and *Staphylococcus aureus*) [22], and one general model to classify contigs of many other species (n=127) that are not frequently represented in databases. All models included predictions from Platon [23] and RFPlasmid[24] which were designed to predict plasmids of multiple species. Mlplasmids was selected as a

third-classifier for *E. faecium*, while novel Centrifuge-based classifiers were developed for the rest of the species. Centrifuge [25] serves as a taxonomy classifier for metagenomics reads, but we adapted this tool to function as a binary classifier of whole genome sequencing (WGS) contigs, similarly to what has been described for PlaScope [26]. In contrast to PlaScope, our Centrifuge classifiers calculate the proportion of hits from each contig in relation to different fractions of the database in order to generate the classification. For example, if a contig matches more than 70% of the times to the plasmidome fraction of the database, it is classified as a ‘plasmid’ contig. In contrast, if it matches less than 30%, it is classified as a ‘chromosome’ contig. Matches ranging from 30 to 70% are labelled ‘unclassified’.

PlasmidEC was the best performing tool for *E. faecium* and *S. enterica*, while for *K. pneumoniae* and *S. aureus*, Centrifuge and plasmidEC performed comparably well, with Centrifuge presenting a higher recall for *K. pneumoniae* and plasmidEC for *S. aureus*. Notably, when classifying genomes of 66 species, less frequently represented in databases, the F1-Scores(contig) of Centrifuge (0.88) and plasmidEC (0.87) surpassed those of RFPlasmid (0.73) and Platon (0.75). Strikingly, despite being a reference based approach, Centrifuge’s general model outperformed other binary classification tools when applied to species not frequently represented in databases. It has been previously observed that plasmids show similarities within genera and even within phylogenetic classes [27]. Possibly, the comprehensive database that we developed with Centrifuge could capture sufficient plasmidome diversity to correctly identify plasmid sequences from novel species, especially if these species are included within frequently represented genera or families. On the other hand, both the Centrifuge database and its benchmark dataset contained genomes from the same public database, which is biased towards clinical isolates from high-income countries [22]. This approach may have positively impacted our results, and the generalizability of the tool’s performance to non-clinical isolates or isolate from low-income countries still has to be determined.

We combined gplas with the best available binary classifier (plasmidEC or centrifuge) and reconstructed individual plasmids of more than 75 species, and compared its performance against MOB-suite and plasmidSPAdes. Gplas demonstrated a good performance when reconstructing AMR-plasmids of *E. faecium*, *K. pneumoniae*, *S. enterica* and *S. aureus*, while MOB-suite and plasmidSPAdes struggled to predict *K. pneumoniae* and *E. faecium* plasmids. We found that isolates from these two species frequently carried a larger number of plasmids and contained more repeats per plasmid, which would lead to more fragmented and entangled assemblies. Furthermore, our results suggest that these plasmidome features (number of plasmids per genome and number of repeats per plasmid) negatively impact the predictions of MOB-suite and plasmidSPAdes. When reconstructing AMR-plasmids of infrequent species, gplas predictions had a considerably higher F1-Score (median=0.85) than MOB-suite (median=0.66) and plasmidSPAdes (median=0.68). Together, these observations suggest that gplas can be successfully applied to reconstruct AMR-plasmids of any species, and that its performance is affected to a lesser extent by variations in the plasmidome features.

### **Future perspectives: Additional applications and further improvements of plasmid prediction and reconstruction tools**

We envision that gplas in combination with plasmidEC’s *general* model could potentially also be applied to metagenomes. However, since metagenomes are considerably more complex than WGS

assembly graphs, several biological constraints and computational challenges will need to be addressed to make gplas feasible: (1) To reduce the run time, plasmid walks will need to be parallelized and the maximum number of nodes explored in each walk would need to decrease; (2) since metagenomes contain sequences from multiple genomes with different abundances, the variation in sequencing coverage of chromosome-derived nodes will not be a useful metric for building plasmid-walks with homogeneous coverage, instead adding taxonomy information to the nodes, and evaluating the coverage differences between pairs of flanking unitigs of the same species could prove useful for estimating this sequencing coverage variation.

The methods developed in this thesis, building upon previously existing plasmid prediction tools, improved the ability to reconstruct plasmids using short-read sequencing data. Nevertheless, we should mention a number of limitations that still need to be addressed. First, plasmidEC requires more computational resources and has longer run times when compared to individual binary classification tools. Although run times could be reduced by multi-threading the predictions of each tool, memory requirements cannot be modified. To address this aspect, we also integrated a *quick mode* into plasmidEC, in which contigs get classified only using Centrifuge (or PlaScope). PlaScope was the fastest binary classification tool and required less memory than most tools when applied to *E. coli*, as described in **chapter 3**. Moreover, the combination of gplas with PlaScope (gplas\_PlaScope) led to individual plasmid predictions with a comparable quality to those obtained with gplas\_plasmidEC. Regarding computational efficiency, it must be noted that Centrifuge loads its database into memory before making predictions, which is not a concern for the species-specific models that we developed. However, the *general* model relies on a database of 30 gigabytes of size, which could limit the possibility of running this model in a standard desktop computer. Second, as most reference-based approaches, Centrifuge and PlaScope cannot classify novel sequences that are not present in their underlying databases. As previously mentioned, public databases mostly contain clinical isolates from high-income countries [22], therefore the application of a strictly reference-based approach could become problematic when attempting to identify plasmid contigs from uncommon niches or geographical locations. An ensemble classifier, such as plasmidEC, that incorporates tools with different computational approaches could potentially be more flexible to predict the origin of novel contigs. Third, mobile genetic elements (MGE) (i.e. transposons, integrative and conjugative elements, integrons) can be carried by both plasmids and chromosomes [17,23,28] and their binary classification without exploration of their flanking sequences, could frequently lead to errors. The propagation of the initial labels (plasmid/chromosome) through the assembly graph, as successfully implemented by plASgraph [30], could improve classification of these elements. A similar approach is implemented the tool GraphBin2 [30], which refines binning results of metagenomics samples.

### Applying plasmidEC and gplas to study the genomic composition of *E. coli* isolates in ICU patients

Despite the limitations discussed above, plasmidEC and gplas were fundamental to performing an in-depth evaluation of the effects of selective digestive decontamination (SDD) on the pangenome composition of extended spectrum beta-lactamase (ESBL) producing *E. coli*. Since it has been suggested that SDD alters the gut microbiome composition of ICU patients [18,19], in **chapter 5**, we hypothesized that these changes in gut ecology might also impact the pangenome composition of potentially pathogenic microorganisms (PPMOs) that continue to populate the intestinal tract despite the use of SDD. To verify this hypothesis, we performed a comparative genomic analysis on

129 ESBL *E. coli* isolates from ICU patients that received either SDD (n=63) or only standard care (n=66). One of the main strengths of our study is its multi-center nature, with isolates obtained in the setting of a cluster-randomized trial from five different ICUs located in three European countries.

The results from our study suggested that SDD does not have a significant impact in shaping the overall pangenome composition of ESBL *E. coli*. Furthermore, the use of plasmidEC allowed us to show that plasmidome composition was not affected by the use of SDD either. Instead, the interplay between phylogeny and geographical location of the ICU explained the largest fraction of plasmidome variance observed between the 129 isolates. Previous studies have described that particular clones and plasmids can persist over time in ICUs and spread to multiple patients [31–35]. We also observed minor differences between the resistomes of SDD and non-SDD isolates. Interestingly, a potential MGE composed of a tobramycin resistance gene and two additional ARGs flanked by IS26 elements, which we termed Tn(TobraR), was more often detected in SDD isolates (n=14, 22%) than in non-SDD isolates (n=4, 6%). The combination of gplas with mge-cluster [36] was essential for determining that the Tn(TobraR) element was found in five different plasmid backbones and 4 different phylogroups (B2, A, C and F), providing further evidence of this element's mobility. Notably, this element frequently co-occurred in the same plasmid with *bla*<sub>CTX-M-15</sub> (an ESBL gene). The existence of Tn(TobraR) and its co-occurrence with *bla*<sub>CTX-M-15</sub> have been previously described in other studies that included *E. coli* isolates from UK, Canada and China [37–39]. It should be noted that we have not collected data on systemic antibiotics provided to these patients, so differences in the resistomes should be interpreted with care.

The limited effect that SDD had on the pangenome composition of ESBL-*E. coli* was unexpected. Although in our study stool samples to assess the effects of SDD on the microbiome composition were not available, other studies have shown that the effect of SDD on the composition of the gut microbiome is substantial [40,41]. These changes in microbiota composition will probably lead to altered gut microbial networks, posing new metabolic challenges to colonising ESBL *E. coli*. It is possible that the adaptation of *E. coli* to the new gut ecology induced by SDD, might have been mediated by changes in gene expression patterns, rather than by loss or acquisition of new genes [42]. Moreover, a recent study demonstrated that *E. coli* auxotrophies can be rescued by expressing short peptides that are coded in novel small open reading frames [43]. Both of these adaptation strategies cannot be detected by the analysis we did in **chapter 5**, which is solely based on gene/presence absence comparisons. It is also possible that the duration of SDD treatment (median time from start of SDD to sample collection of the sequenced ESBL *E. coli* isolate 4, IQR = 2 - 6.5 days) might have not been sufficiently long to cause an appreciable change in the community structure of PPMOs. Long-term follow-up with sequencing of multiple and/or repeat isolates would have allowed to demonstrate changes in the ESBL *E. coli* population or specific isolates. Furthermore, the use of ESBL-selective media for surveillance cultures in the primary study precluded analysis of the entire *E. coli* population in these patients, for which adaptations in non-ESBL *E. coli* isolates may have been missed. This selection could have camouflaged changes in the genome composition of different *E. coli* subpopulations within each patient.

**Future clinical directions: towards affordable and accurate plasmid surveillance systems**

Given their superior resolution power, whole genome sequencing (WGS) approaches are becoming the new gold-standard for assessing potential outbreaks of resistant pathogens in clinical settings [44–50]. Although pathogen clonality is often the main question, limiting surveillance only to bacterial clones provides an incomplete picture of the risks of AMR dissemination. Essentially this approach precludes from detecting outbreaks mediated by MGE. Recently, plasmid-mediated multispecies outbreaks have been detected in single hospitals [51–53] and also at a country-wide scale [54]. Consequently, routine genomic surveillance of plasmids and other MGEs, provides valuable extra information for an early detection of outbreaks of AMR bacteria. Although complete reconstruction of plasmids generally requires long-read sequencing, at the moment it is still prohibitively expensive for routine application (especially surveillance), despite the fact that WGS surveillance can be intentionally limited to select populations of vulnerable patients (such as those admitted into ICUs). Consequently, the methods that we developed in **chapter 3** and **4** of this thesis, which provide accurate reconstruction of plasmids using affordable short reads, could permit the implementation of regular WGS-based plasmid surveillance in clinical settings, facilitating the early detection of plasmid-mediated outbreaks, and ultimately guiding infection prevention initiatives.

Additionally, routine surveillance systems at a national or international level are important to explore the trends of resistance in different regions and niches [55,56] and also to detect and interrupt (multi-) country-wide outbreaks that might lead to the dissemination of AMR [54, 57]. In 2016, the European Center for Disease Prevention and Control (ECDC) claimed that WGS-based typing should become the primary microbial typing method for the investigation of multi-country outbreaks and antimicrobial surveillance in the European Union, at least for bacterial pathogens [58]. Since then, WGS capacity in reference laboratories has increased across different countries, an EU-wide data sharing platform was built, and harmonised WGS-based bacterial typing methods have been defined [59]. The incorporation of affordable methods to reconstruct plasmid sequences into these large-scale international surveillance initiatives could aid in detecting the dissemination of particularly dangerous resistant plasmids or plasmid-clone combinations. In turn, this information could guide the development and implementation of more-affordable and faster surveillance methodologies in clinical settings, for example, via the design of multiplex PCR schemes targeting specific plasmid modules and resistance determinants simultaneously.

**Final words**

In this thesis we improved existing methods to reconstruct plasmids from short-read data. We described how feature-enriched assembly graphs, as implemented in *gplas*, contribute to generating predictions that include higher fractions of AMR-plasmids across multiple species, given that an accurate binary classifier exists. Moreover, we demonstrated that it is possible to use these plasmid predictions to track smaller genomic elements, such as the *Tn*(*TobRaR*), across multiple plasmid backbones. We hope that coupling these methods to novel tools to type plasmids, such as *mge-cluster*, will make epidemiological plasmid studies more accessible and widespread, and can ultimately be used for WGS-based surveillance of AMR using short-reads.

## References

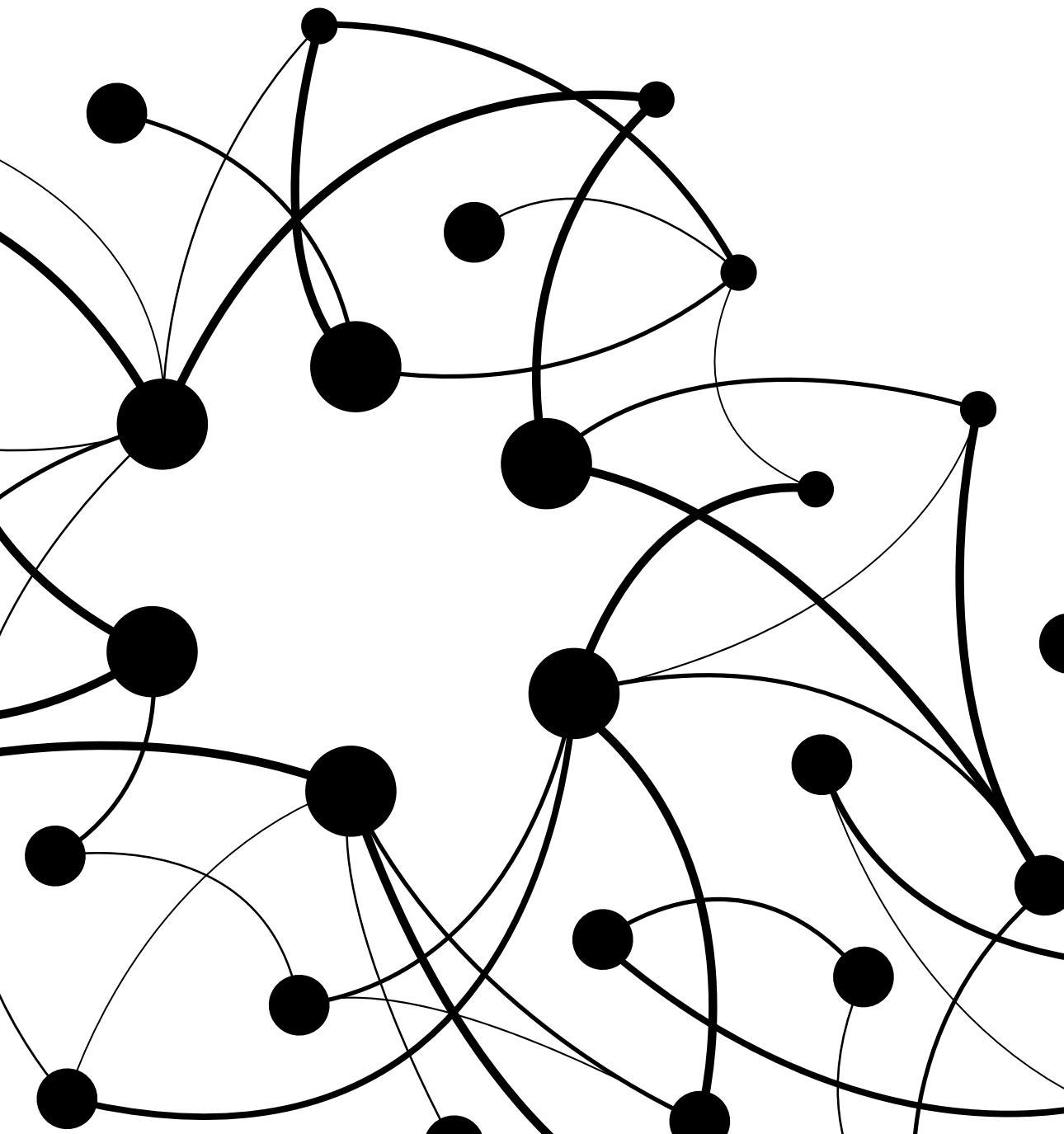
1. Mane A, Faizrahneem M, Chauve C. A Mixed Integer Linear Programming Algorithm for Plasmid Binning. *Comparative Genomics*. 2022; 279–292.
2. Pu L, Shamir R. 4CAC: 4-class classification of metagenome assemblies using machine learning and assembly graphs. *bioRxiv*. 2023. p. 2023.01.20.524935. doi:10.1101/2023.01.20.524935
3. Pu L, Shamir R. 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics*. 2022;38: ii56–ii61.
4. Gomi R, Wýres KL, Holt KE. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microbial genomics*. 2021;7. doi:10.1099/mgen.0.000550
5. 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.
6. Ebmeyer S, Kristiansson E, Larsson DGJ. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Commun Biol*. 2021;4: 8.
7. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. 2022;399: 629–655.
8. Kawamura K, Nagano N, Suzuki M, Wachino J-I, Kimura K, Arakawa Y. ESBL-producing and Its Rapid Rise among Healthy People. *Food Saf (Tokyo)*. 2017;5: 122–150.
9. Matamoros S, van Hattem JM, Arcilla MS, Willems N, Melles DC, Penders J, et al. Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the *mcr-1* gene indicates bacterial diversity but plasmid restriction. *Sci Rep*. 2017;7: 1–9.
10. Zamudio R, Boerlin P, Beyrouthy R, Madec J-Y, Schwarz S, Mulvey MR, et al. Dynamics of extended-spectrum cephalosporin resistance genes in *Escherichia coli* from Europe and North America. *Nat Commun*. 2022;13: 7490.
11. Pires J, Huber L, Hickman RA, Dellicour S, Lunha K, Leangapichart T, et al. Genome-associations of extended-spectrum  $\beta$ -lactamase producing (ESBL) or AmpC producing *E. coli* in small and medium pig farms from Khon Kaen province, Thailand. *BMC Microbiol*. 2022;22: 253.
12. Huisman JS, Vaughan TG, Egli A, Tschudin-Sutter S, Stadler T, Bonhoeffer S. The effect of sequencing and assembly on the inference of horizontal gene transfer on chromosomal and plasmid phylogenies. *Philos Trans R Soc Lond B Biol Sci*. 2022;377: 20210245.
13. Nguyen M, Olson R, Shukla M, VanOeffelen M, Davis JJ. Predicting antimicrobial resistance using conserved genes. *PLoS Comput Biol*. 2020;16: e1008319.
14. Arredondo-Alonso S, Top J, Corander J, Willems RJL, Schürch AC. Mode and dynamics of vanA-type vancomycin resistance dissemination in Dutch hospitals. *Genome Med*. 2021;13. doi:10.1186/s13073-020-00825-3
15. Venturini C, Hassan KA, Chowdhury PR, Paulsen IT, Walker MJ, Djordjevic SP. Sequences of Two Related Multiple Antibiotic Resistance Virulence Plasmids Sharing a Unique IS26-Related Molecular Signature Isolated from Different *Escherichia coli* Pathotypes from Different Hosts. *PLoS One*. 2013;8: e78862.
16. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Science advances*. 2021;7. doi:10.1126/sciadv.abe3868
17. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, 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. doi:10.1073/pnas.2008731118
18. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016;32: 3380–3387.
19. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, et al. Recycler: an algorithm for



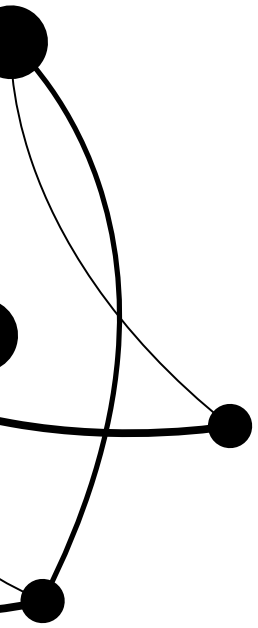
- detecting plasmids from de novo assembly graphs. *Bioinformatics*. 2017;33: 475–482.
20. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. *gplas*: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics*. 2020;36: 3874–3876.
  21. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. *mplasmids*: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000224
  22. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol*. 2021;19: e3001421.
  23. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. *Platon*: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom*. 2020;6. doi:10.1099/mgen.0.000398
  24. 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. doi:10.1099/mgen.0.000683
  25. Kim D, Song L, Breitwieser FP, Salzberg SL. *Centrifuge*: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 2016;26: 1721–1729.
  26. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. *PlaScope*: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom*. 2018;4. doi:10.1099/mgen.0.000211
  27. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun*. 2020;11: 1–13.
  28. Wang Y, Batra A, Schulenburg H, Dagan T. Gene sharing among plasmids and chromosomes reveals barriers for antibiotic resistance gene transfer. *Philos Trans R Soc Lond B Biol Sci*. 2022;377: 20200467.
  29. Sielemann J, Sielemann K, Brejová B, Vinař T, Chauve C. *plASgraph* - using graph neural networks to detect plasmid contigs from an assembly graph. *bioRxiv*. 2022. p. 2022.05.24.493339. doi:10.1101/2022.05.24.493339
  30. Mallawaarachchi VG, Wickramarachchi AS, Lin Y. Improving metagenomic binning results with overlapped bins using assembly graphs. *Algorithms Mol Biol*. 2021;16: 3.
  31. Klassert TE, Leistner R, Zubiria-Barrera C, Stock M, López M, Neubert R, et al. Bacterial colonization dynamics and antibiotic resistance gene dissemination in the hospital environment after first patient occupancy: a longitudinal metagenetic study. *Microbiome*. 2021;9: 1–17.
  32. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. 2020 [cited 21 Feb 2023]. doi:10.7554/eLife.53886
  33. Brodrick HJ, Raven KE, Kallonen T, Jamrozny D, Blane B, Brown NM, et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Med*. 2017;9: 1–11.
  34. Hawkey J, Wyres KL, Judd LM, Harshegyi T, Blakeway L, Wick RR, et al. ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Med*. 2022;14: 1–13.
  35. Hu Y, Zhang H, Wei L, Feng Y, Wen H, Li J, et al. Competitive Transmission of Carbapenem-Resistant *Klebsiella pneumoniae* in a Newly Opened Intensive Care Unit. *mSystems*. 2022;7. doi:10.1128/mSystems.00799-22
  36. Arredondo-Alonso S, Gladstone RA, Pöntinen AK, Gama JA, Schürch AC, Lanza VF, et al. Consistent typing of plasmids with the *mge*-cluster pipeline. *bioRxiv*. 2022. p. 2022.12.16.520696. doi:10.1101/2022.12.16.520696
  37. Livermore DM, Day M, Cleary P, Hopkins KL, Toleman MA, Wareham DW, et al. OXA-1  $\beta$ -lactamase and non-susceptibility to penicillin/ $\beta$ -lactamase inhibitor combinations among ESBL-producing *Escherichia coli*. *J Antimicrob Chemother*. 2019;74: 326–333.

38. Peirano G, Pitout JDD. Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4. *Int J Antimicrob Agents*. 2010;35: 316–321.
39. Cao X, Zhang Z, Shen H, Ning M, Chen J, Wei H, et al. Genotypic characteristics of multidrug-resistant *Escherichia coli* isolates associated with urinary tract infections. *APMIS*. 2014;122: 1088–1095.
40. Benus RF, Harmsen HJ, Welling GW, Spanjersberg R, Zijlstra JG, Degener JE, et al. Impact of digestive and oropharyngeal decontamination on the intestinal microbiota in ICU patients. *Intensive Care Med*. 2010;36. doi:10.1007/s00134-010-1826-4
41. van Doorn-Schepens MLM, Abis GSA, Oosterling SJ, van Egmond M, Poort L, Hbac S, et al. The effect of selective decontamination on the intestinal microbiota as measured with IS-pro: a taxonomic classification tool applicable for direct evaluation of intestinal microbiota in clinical routine. *Eur J Clin Microbiol Infect Dis*. 2022;41. doi:10.1007/s10096-022-04483-8
42. Gagarinova A, Hosseinnia A, Rahmatbakhsh M, Istace Z, Phanse S, Moutaoufik MT, et al. Auxotrophic and prototrophic conditional genetic networks reveal the rewiring of transcription factors in *Escherichia coli*. *Nat Commun*. 2022;13: 1–16.
43. Babina AM, Surkov S, Ye W, Jerlström-Hultqvist J, Larsson M, Holmqvist E, et al. Rescue of *Escherichia coli* auxotrophy by de novo small proteins. 2023 [cited 21 Mar 2023]. doi:10.7554/eLife.78299
44. Li L, Wang R, Qiao D, Zhou M, Jin P. Tracking the Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* in an Emergency Intensive Care Unit by Whole Genome Sequencing. *Infect Drug Resist*. 2022;15. doi:10.2147/IDR.S386385
45. Egan SA, Corcoran S, McDermott H, Fitzpatrick M, Hoyne A, McCormack O, et al. Hospital outbreak of linezolid-resistant and vancomycin-resistant ST80 *Enterococcus faecium* harbouring an *optrA*-encoding conjugative plasmid investigated by whole-genome sequencing. *J Hosp Infect*. 2020;105. doi:10.1016/j.jhin.2020.05.013
46. Rubin LG, Beachy J, Matz T, Balamohan A, Jendresky L, Zembera J, et al. Prolonged outbreak of clonal, mupirocin-resistant methicillin-resistant *Staphylococcus aureus* in a neonatal intensive care unit: association with personnel and a possible environmental reservoir, analyzed using whole genome sequencing. *Am J Infect Control*. 2022;50. doi:10.1016/j.ajic.2021.09.010
47. Tang M, Li J, Liu Z, Xia F, Min C, Hu Y, et al. Clonal transmission of polymyxin B-resistant hypervirulent *Klebsiella pneumoniae* isolates coharboring blaNDM-1 and blaKPC-2 in a tertiary hospital in China. *BMC Microbiol*. 2023;23. doi:10.1186/s12866-023-02808-x
48. Brizuela J, Kajeekul R, Roodsant TJ, Riwoad A, Boueroy P, Pattanapongpaibool A, et al. *Streptococcus suis* outbreak caused by an emerging zoonotic strain with acquired multi-drug resistance in Thailand. *Microbial genomics*. 2023;9. doi:10.1099/mgen.0.000952
49. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin Microbiol Rev*. 2017;30: 1015–1063.
50. Clinical perspectives in integrating whole-genome sequencing into the investigation of healthcare and public health outbreaks – hype or help? *J Hosp Infect*. 2021;109: 1–9.
51. Yamagishi T, Matsui M, Sekizuka T, Ito H, Fukusumi M, Uehira T, et al. A prolonged multispecies outbreak of IMP-6 carbapenemase-producing Enterobacterales due to horizontal transmission of the IncN plasmid. *Sci Rep*. 2020;10: 4139.
52. Mari-Almirall M, Ferrando N, Fernández MJ, Cosgaya C, Viñes J, Rubio E, et al. Clonal Spread and Intra- and Inter-Species Plasmid Dissemination Associated With *Klebsiella pneumoniae* Carbapenemase-Producing Enterobacterales During a Hospital Outbreak in Barcelona, Spain. *Front Microbiol*. 2021;12. doi:10.3389/fmicb.2021.781127

53. Hidalgo L, de Been M, Rogers MRC, Schürch AC, Scharringa J, van der Zee A, et al. Sequence-based epidemiology of an OXA-48 plasmid during a hospital outbreak. *Antimicrob Agents Chemother.* 2019;63. doi:10.1128/AAC.01204-19
54. European Centre for Disease Prevention and Control. Combined clonal and plasmid-mediated outbreak of carbapenemase-producing Enterobacterales in Lithuania, 2019-2020 – 3 February 2020. ECDC: Stockholm; 2020.
55. European Food Safety Authority (EFSA), European Centre for Disease Prevention and Control (ECDC). The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2020/2021. *EFSA J.* 2023;21: e07867.
56. European Centre for Disease Prevention and Control. Antimicrobial resistance in the EU/EEA (EARS-Net) - Annual Epidemiological Report 2021. Stockholm: ECDC; 2022.
57. Rapid risk assessment: Carbapenemase-producing (OXA-48) *Klebsiella pneumoniae* ST392 in travellers previously hospitalised in Gran Canaria, Spain. In: European Centre for Disease Prevention and Control [Internet]. 11 Jul 2018 [cited 14 Mar 2023]. Available: <https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-carbapenemase-producing-oxa-48-klebsiella-pneumoniae-st392>
58. Expert opinion on whole genome sequencing for public health surveillance. In: European Centre for Disease Prevention and Control [Internet]. 15 Aug 2016 [cited 14 Mar 2023]. Available: <https://www.ecdc.europa.eu/en/publications-data/expert-opinion-whole-genome-sequencing-public-health-surveillance>
59. ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations. In: European Centre for Disease Prevention and Control [Internet]. 4 Apr 2019 [cited 14 Mar 2023]. Available: <https://www.ecdc.europa.eu/en/publications-data/ecdc-strategic-framework-integration-molecular-and-genomic-typing-european>



# Appendices



### English Summary

Antimicrobial resistance (AMR) is a pressing global concern, posing a significant threat to human health worldwide. Each year, the incidence of infections caused by resistant bacteria continues to rise, imposing a significant challenge for healthcare professionals. The emergence of multidrug resistant (MDR) strains further exacerbates the issue, severely limiting therapeutic options available for managing infections. Adding to the complexity of the problem, healthcare facilities experience a higher prevalence of MDR bacterial infections, primarily due to the extensive and frequent utilization of antibiotics in these settings.

This thesis is mainly focused on *Escherichia coli*, a bacterium that commonly resides as a commensal in the gastrointestinal tract of humans and other warm-blooded animals. Nevertheless, over the past decades, *E. coli* has become one of the most prevalent MDR pathogens worldwide. In particular, infections caused by extended spectrum beta-lactamase (ESBL)-producing *E. coli* have rapidly increased and have become an important public health concern.

AMR determinants, such as ESBL genes, are frequently encoded by plasmids, mobile genetic elements (MGE) that can be horizontally disseminated across bacteria by diverse mechanisms. Multiple studies suggest that plasmids are essential drivers of the spread of resistance within diverse ecological niches. Additionally, plasmids also play a crucial role in the development of outbreaks in clinical settings that involve multiple bacterial species. Consequently, accurate plasmid identification and tracking are necessary to better understand the mechanisms that drive AMR dissemination.

The development of next-generation sequencing (NGS) platforms has allowed high-throughput bacterial genome research. Illumina short reads remain the most widespread sequencing technology worldwide. These platforms produce highly accurate reads and large amounts of samples can be processed simultaneously. Nevertheless, due to the frequent occurrence of large repeated elements in bacterial genomes, the *de novo* assembly of short reads produces hundreds of contigs of unclear origin (plasmid or chromosome) mingled together in a draft genome. Consequently, determining the exact sequence of plasmids using short reads alone is challenging.

In **chapter 2** of this thesis, we conducted a comprehensive review of various bioinformatic tools designed to predict plasmid sequences from short-read sequencing data. Furthermore, we extensively compared the performance of six plasmid reconstruction tools using a dataset comprising 240 publicly available *E. coli* genomes from databases. Through this benchmark study, we identified MOB-suite as the most accurate tool for reconstructing *E. coli* plasmids. Nonetheless, the evaluation also revealed that all tools encountered challenges in reconstructing large AMR plasmids.

In **chapter 3**, recognizing the limitations of existing plasmid reconstruction tools, we devised a novel two-step approach to reconstruct *E. coli* plasmids from Illumina sequencing data. In the initial step, we employed plasmidEC, an ensemble classifier developed by integrating three distinct binary classification tools, to classify nodes from the assembly graph as plasmid- or chromosome-derived. Subsequently, we applied gplas to bin plasmid-derived nodes into individual plasmid predictions based on sequencing coverage similarities and assembly graph connectivity. Additionally, gplas was adapted to yield improved reconstructions of plasmids exhibiting large sequencing coverage variations. Our method outperformed MOB-suite, particularly when reconstructing AMR plasmids.

In **chapter 4**, we extended the application of our plasmid reconstruction method to multiple species. To achieve this, we developed four species-specific plasmidEC models for *E. faecium*, *K. pneumoniae*, *S. enterica*, and *S. aureus*, alongside one species-independent model for less-frequent species. PlasmidEC presented the best performance when identifying plasmid contigs from *E. faecium*, *S. enterica*, and *S. aureus*, while Centrifuge outperformed all other tools for *K. pneumoniae* and less-frequent species. Through the integration of the best-performing binary classifiers with gplas, we were able to reconstruct plasmids of diverse species. We evaluated the tool's performance by reconstructing 953 plasmids from 70 different bacterial species and compared the results to MOB-suite and plasmidSPAdes reconstructions. Remarkably, gplas displayed a consistent performance when reconstructing large AMR plasmids across multiple species, while the other tools exhibited greater variations in performance.

In **chapter 5**, we compared the pangenomes, plasmidomes and resistome compositions of 129 ESBL-*E. coli* isolates, obtained from patients included in the R-GNOSIS ICU study. In this cluster-randomized crossover trial, patients admitted to intensive care units (ICU) did or did not receive selective digestive decontamination (SDD) as a prophylactic treatment to prevent colonization with potentially pathogenic microorganisms. SDD consists of a mix of topical antibiotics (Tobramycin, Colistin and Amphotericin B) that target aerobic gram-negative bacteria, *P. aeruginosa*, *S. aureus* and yeast, but do not compromise the anaerobic flora. The data analyzed in this chapter included ESBL-*E. coli* isolates from five different ICUs located in Spain, Belgium and the UK. Our findings suggested that SDD has a limited impact on the population structure and pangenome composition of ESBL-*E. coli*. However, isolates obtained from patients that received standard care had a higher amount of aminoglycoside resistance genes, while SDD isolates were more frequently found to possess a transposon carrying a tobramycin resistance gene. This transposon contained a total of three AMR genes surrounded by IS26 elements, and frequently co-occurred with *bla*<sub>CTX-M-15</sub> in multiple clones and distinct plasmid backbones.

In this thesis I explored the challenges of existing bioinformatic tools when attempting to reconstruct AMR plasmids from short-read sequencing data. Furthermore, I also described a novel two-step approach to reconstruct these plasmids. This novel approach consists of combining an accurate binary classifier of contigs with gplas, a tool that bins plasmid-derived contigs into individual plasmid predictions based on coverage uniformity and assembly graph connectivity. Our novel method proved to be accurate when reconstructing plasmids of more than 70 bacterial species. Finally, I studied the epidemiology of ESBL-*E. coli* isolates and plasmids obtained from different ICUs across Europe.

### Nederlandse samenvatting

Antimicrobiële resistentie (AMR) is een urgent, wereldwijd probleem en vormt een aanzienlijke bedreiging voor de menselijke gezondheid. Elk jaar neemt het aantal infecties veroorzaakt door resistente bacteriën toe, wat een grote uitdaging vormt voor de gezondheidszorg. De opkomst van multidrug-resistente (MDR)-stammen verergert het probleem verder, waardoor de beschikbare therapeutische opties voor het beheersen van infecties verder worden beperkt. Wat het probleem nog ingewikkelder maakt, is dat zorginstellingen een hogere prevalentie van MDR-bacteriële infecties kennen, voornamelijk als gevolg van het veelvuldige gebruik van antibiotica in deze omgevingen.

Dit proefschrift is voornamelijk gericht op *Escherichia coli*, een bacterie die gewoonlijk als commensaal aanwezig is in het maag-darmkanaal van mensen en andere warmbloedige dieren. Niettemin is *E. coli* de afgelopen decennia een van de meest voorkomende MDR-pathogenen ter wereld geworden. Met name infecties veroorzaakt door *E. coli* die een uitgebreid spectrum bèta-lactamase (ESBL) produceren, zijn snel toegenomen en zijn een belangrijk probleem voor de volksgezondheid geworden.

AMR-determinanten, zoals ESBL-genen, worden vaak gecodeerd door plasmiden, mobiele genetische elementen (MGE) die via verschillende mechanismen horizontaal tussen bacteriën kunnen worden verspreid. Meerdere studies suggereren dat plasmiden essentiële aanjagers zijn van de verspreiding van resistentie binnen diverse ecologische niches. Bovendien spelen plasmiden ook een cruciale rol bij de totstandkoming van uitbraken in klinische omgevingen waarbij meerdere bacteriesoorten betrokken zijn. Hierdoor is nauwkeurige identificatie en tracking van plasmiden nodig om de mechanismen die de verspreiding van AMR faciliteren beter te begrijpen.

De ontwikkeling van next-generation sequencing (NGS) platforms heeft high-throughput bacterieel genoomonderzoek mogelijk gemaakt. 'Illumina short reads' blijft wereldwijd de meest wijdverbreide sequencing-technologie. Deze platforms produceren zeer nauwkeurige reads en grote hoeveelheden monsters kunnen tegelijkertijd worden verwerkt. Echter, als gevolg van het veelvuldig voorkomen van grote repetitieve elementen in bacteriële genomen, produceert de de novo assembly van korte reads honderden contigs van onduidelijke oorsprong (plasmide of chromosoom) die met elkaar vermengd zijn in een conceptgenoom. Hierdoor is het een uitdaging om de exacte sequentie van plasmiden te bepalen met alleen korte reads.

In **hoofdstuk 2** van dit proefschrift hebben we een uitgebreid overzicht gegeven van verschillende bio-informatica-tools die zijn ontworpen om plasmide-sequenties te voorspellen op basis van short-read sequencing data. Verder hebben we de nauwkeurigheid van zes plasmide-reconstructie-tools uitgebreid vergeleken met behulp van een dataset bestaande uit 240 openbaar beschikbare *E. coli*-genomen uit databases. Door middel van deze benchmarkstudie hebben we MOB-suite aangewezen als de meest nauwkeurige tool voor het reconstrueren van *E. coli*-plasmiden. Desalniettemin onthulde de evaluatie ook dat alle tools problemen ondervonden bij het reconstrueren van grote AMR-plasmiden.

In **hoofdstuk 3**, waarin we de beperkingen van bestaande hulpmiddelen voor plasmide reconstructie onderkennen, hebben we een nieuwe tweetrapsbenadering ontwikkeld om *E. coli*-plasmiden te reconstrueren uit Illumina-sequencing data. In de eerste stap hebben we plasmidEC gebruikt, een ensemble-classificatietool die is ontwikkeld door drie verschillende binaire classificatietools te



integreren, om punten uit de assembly graph te classificeren als zijnde afkomstig van een plasmide of een chromosoom. Vervolgens hebben we gplas toegepast om punten afkomstig van plasmiden in individuele plasmide voorspellingen te plaatsen op basis van overeenkomsten in sequence coverage en connectiviteit van assembly graphs. Bovendien werd gplas aangepast om verbeterde reconstructies van plasmiden op te leveren die grote variaties in de sequence coverage vertoonden. Onze methode presteerde beter dan MOB-suite, vooral bij het reconstrueren van AMR-plasmiden.

In **hoofdstuk 4** hebben we de toepassing van onze plasmide-reconstruatiemethode uitgebreid naar meerdere soorten. Hiertoe hebben we vier soortspecifieke modellen ontwikkeld voor *E. faecium*, *K. pneumoniae*, *S. enterica* en *S. aureus*, en een soort-onafhankelijk plasmidEC-model. PlasmidEC presteerde het beste bij het identificeren van plasmide-contigs van *E. faecium*, *S. enterica* en *S. aureus*, terwijl Centrifuge beter presteerde dan alle andere tools voor *K. pneumoniae* en minder frequente soorten. Door de integratie van de best presterende binaire classificers met gplas, waren we in staat om plasmiden van diverse soorten te reconstrueren. We hebben de prestaties van de tool geëvalueerd door 953 plasmiden van 70 verschillende bacteriesoorten te reconstrueren en de resultaten te vergelijken met reconstructies van MOB-suite en plasmidSPAdes. Opmerkelijk genoeg vertoonde gplas een consistente prestatie bij het reconstrueren van grote AMR-plasmiden over meerdere soorten, terwijl de andere tools grotere variaties in prestaties vertoonden.

In **hoofdstuk 5** hebben we de pan-genomen, plasmiden en resistoom-samenstellingen vergeleken van 129 ESBL-*E. coli*-isolaten, verkregen van patiënten die deelnamen aan de R-GNOSIS ICU-studie. In deze cluster-gerandomiseerde cross-over studie kregen patiënten die werden opgenomen op de intensive care (ICU) al dan niet selectieve digestieve decontaminatie (SDD) als profylactische behandeling om kolonisatie met potentieel pathogene micro-organismen te voorkomen. SDD bestaat uit een mix van lokale antibiotica (tobramycine, colistine en amfotericine B) die zich richten op aerobe gramnegatieve bacteriën, *P. aeruginosa*, *S. aureus* en gist, maar de anaerobe flora niet in gevaar brengen. De in dit hoofdstuk geanalyseerde data omvat ESBL-*E. coli*-isolaten van vijf verschillende ICU's in Spanje, België en het VK. Onze bevindingen suggereerden dat SDD een beperkte invloed heeft op de populatiestructuur en de samenstelling van het pan-genoom van ESBL-*E. coli*. Isolaten verkregen van patiënten die standaardzorg kregen, hadden echter een grotere hoeveelheid resistentiegenen die zorgen voor resistentie tegen aminoglycoside, terwijl SDD-isolaten vaker een transposon bleken te bezitten dat een tobramycine resistentiegen bevatte. Dit transposon bevatte in totaal drie AMR-genen omgeven door IS26-elementen en kwam vaak samen met *bla*<sub>CTX-M-15</sub> voor in meerdere stammen en verschillende plasmide-backbones.

In dit proefschrift heb ik de uitdagingen onderzocht van bestaande bioinformatica-tools bij het reconstrueren van AMR-plasmiden uit short-read sequencing data. Verder beschrijf ik ook een nieuwe tweetrapsbenadering om deze plasmiden te reconstrueren. Deze nieuwe benadering bestaat uit het combineren van een nauwkeurig binaire classificatie-algoritme van contigs met gplas, een tool die van contigs afkomstig van plasmiden in bins onderverdeelt op basis van uniformiteit van sequence coverage en assembly graph connectiviteit. Onze nieuwe methode bleek zeer nauwkeurig te zijn bij het reconstrueren van plasmiden van meer dan 70 bacteriesoorten. Ten slotte heb ik me verdiept in de epidemiologie van ESBL-*E. coli*-isolaten en plasmiden verkregen van verschillende ICU's in heel Europa.

### Acknowledgments

**Anita**, I really couldn't have asked for a better supervisor and mentor! I always felt that you valued my opinions and you gave me freedom to pursue my own research ideas. I want to thank you for trusting me since the very first day I arrived at the UMCU and for all your support during every stage of my PhD. Thanks also for being so understanding, I always felt very secure to come forward to you with any issue (scientific or personal) I might be going through. Probably you already know this, but you have built such an amazing group! I always felt a part of the MMBioIT group, despite the weird conditions in which I started my PhD. You are a great scientist and an amazing person, and I've learnt so much by working with you. If I'm ever in charge of leading a team (in whatever field), I will aim at doing it as good as you!

**Nienke**, I am extremely happy to have had you as my co-promotor. You are really an outstanding researcher, with an endless curiosity and a super brilliant mind! Thanks for putting so much effort into improving all chapters of my thesis, I always found your inputs extremely insightful and I honestly admire your knowledge and attention to detail. You are also a very warm and kind person, I always felt very supported by you and you always brought such a positive energy into every meeting we ever had, thank you for that! I feel I have learnt a lot from you, both at the scientific and personal level.

**Rob**, you were an amazing PhD promotor. I want to thank you for all the time and knowledge you have shared with me, which has been really invaluable for my development as a scientist and as a person. Your questions, inputs and ways of approaching problems were always 'food for thought'. I also wanted to thank you for bringing such a relaxed energy into every meeting we had and for being so accessible; all of these really contributed to me having a great time during my PD. I've always admired the degree of scientific curiosity you have, even after all the years you've been doing science. Moving forward, I will aim at retaining that level of curiosity.

**Jesse**, the Young Wolf of the MMBioIT team! Man, you are a brilliant bioinformatician, it has been a pleasure to collaborate with you and I've learnt a lot while working by your side. More importantly, I'm super happy that I get to call you my friend! I know we'll get together many times in the future. Remember that you'll always be welcomed in our house, no matter when you read this or where I'm living (probably it will be Spain). Also, please remember to arrive half an hour later of the meeting time I suggest!

**Sergi**, el rey de los plásmidos, michi miao, muchas gracias por tu ayuda durante todo mi doctorado, realmente fuiste un supervisor extra y estoy muy agradecido de que hayas compartido tanto tiempo y conocimiento conmigo. Realmente sos un científico y una persona "de puta madre". Muchas gracias también por tu amistad! Siempre me lo paso muy bien con vos y con **Maria**, y me encanta que nos sigamos encontrando fuera de NL, en Barcelona, en Sitges o donde sea.

**Lisa**, thanks a lot for all the hard work that you put in your Master thesis! PlasmidEC is a big part of my PhD thesis. It was a pleasure working with you, you are a very smart and easy going person. I see nothing but success in your future!

**Matteo**, bambino, fratello, pipiripapi... Thanks for all those interesting and fun discussions, about science but also about life. I'm sure we'll meet in Italy soon and later in Argentina... and then, finally, we'll be able to settle the score on who makes the best Pizza. Espero que hayas aprendido Español para entonces! I'm very glad to have met you and **Ludo**, you guys have really made my (our) stay in the Netherlands much better!

**Malbert**, it has been very nice sharing time with you. You are a great bioinformatician and an excellent guy. You also have a special talent for demolishing old chimneys xD. **Rodrigo** it has been great to collaborate with you in a few projects, you are a very meticulous, hard-working and knowledgeable bioinformatician! You are also an excellent and very warm person. **Wheizen**, it has been a pleasure working with you! I really admire your perseverance and desire to understand things to the deepest of levels. Thanks for the multiple discussions we've had. You are also a very kind person and a very funny guy. **Janetta**, you are truly a remarkable person and scientist. Your level of motivation and love for science is incredible. Thanks for asking me so many great questions during my presentations and for giving me many valuable inputs. Also, thank you for always being so accessible, collaborative and in a good mood! The Willems group is very lucky to have you! **Vicky**, que lindo habernos conocido! Sos una excelente bioinformática, y aun una mejor persona y compañera de equipo. Me has caído fenomenal desde el principio, y como dijiste vos una vez "Parece que nos conocemos desde hace mucho tiempo".

**Paul!** My favourite mongol! Thanks for your warm friendship since day one! I really don't know if I would have survived the pandemic in the NL without you. I'm very happy that you introduced me to Leire, but you will always be my first European Valentine's. We still have a trip to Argentina in our bucket list, I know we'll make it happen!

**Julita** querida! Eres muy majica tia, joder! O en Argentino: Che boluda, re buena piba sos! Muchas gracias por tu buena energía de siempre, Julita, y por aportar tanto al espíritu de grupo! La oficina y el MMB no serían lo mismo sin vos la verdad. "Cariña" y yo te vamos a extrañar mucho. Gracias también (a vos y al loco Wouter) por abrirnos las puertas de su casa y por organizar unas fiestas tan guapas (lindas)!

**Jiannan**, my man, I miss you! I loved sharing the office with you. Thanks for the fun memes, drinks and positivity! "It's fine, it's fine", became our daily mantra!

**Jalal**, thanks for all the fun time together. We both know that Argentina wouldn't have won the world cup without you watching all the matches with me! Hope to visit you soon in Tromso.

**Sjors**, muchas gracias for all the nice meals, airport nights and bad music taste for concerts in far away lands! I wish you and Inge the best, and I have no doubts we will be in close contact!

**Jelle**, the crazy father of Paul and the second husband of Leire! That combination can only make you someone great, fun and special! Thanks for all the great moments! I wish you a great retirement!

For the office guys, **Frerich, Leo, Tristan, Michał, Barat and Coco**, thanks for all the beer and memes moments! It was a lot of fun working with you, you made my PhD days more memorable! Special thanks to brave **Leo**, for always donating his house for doing fun stuff, at the risk of it getting destroyed.

**Marieke, Nine, Lotte, Kulsum, Dani, Bart, Remy, Eva, Lisanne** and the rest and former PhDs and postdocs in our department. It's been a pleasure working with you, thanks for all the time that we spent together!

A la tía **Bettina**, gracias tiuuchiiii por todo tu amor y por aguantarnos tantas veces en USA con máxima paciencia! Muchas gracias también por venir a visitarnos a Holanda, me hizo muy muy bien, en un momento muy duro! Te queremos muuucho!

**Nico** querido, amigo, compañero de estudio, compañero de aventuras y, recientemente, compañero de piso también! ¡Cómo te quiero wacho! ¡Sos realmente una persona del bien! Admiro muchísimo tu rectitud, tu inteligencia, tu curiosidad, tu bondad y tu paciencia! La verdad, no podría haber perdido un mejor amigo. Siento que es una hermosa revancha y un maravilloso honor que puedas estar cuando defienda mi tesis y que encima seas mi paranymp!

A mis mapadres: **Ali y Ozzy**. Gracias por su sacrificio y por su amor incondicional. Siempre hicieron de más: más de lo que debían y más de lo que podían. Mil gracias! Esta tesis nunca hubiera existido si ustedes no se rompían el alma por nosotros. Fueron, son y serán los mejores mapadres que yo pudiera haber querido. ¡Los quiero muchísimo!

A mi querida hermanita **Agus**, gracias por estar siempre al lado mío, por empujar para adelante y por tirar la mejor onda a cada momento. ¡Sos gigante! Gracias también por bancar los trapos sola en Argentina durante estos últimos años. Que aburrida y chota la vida sin hermanos no? Menos mal que estas aca! ¡Te quiero mucho mucho!

**Clau**, no se ni por dónde empezar... ¡Qué bendición tenerte en nuestra vida la verdad! Gracias por tu amor, siempre tan verdadero y desinteresado. Gracias por cuidarnos tanto y por transmitirnos valores tan buenos. Gracias por estar siempre, en las buenas, en las masomenos, en las malas y en las muy malas. Honestamente, no sé dónde estaríamos sin vos. Te quiero infinito Clau.

A la tía **Naná**, como me gustaría que estuvieras acá. Gracias por tu amor y por tu paciencia, ambos inagotables. Gracias por dedicar tanto tiempo y esfuerzo a nuestra educación. Te quiero mucho tía.

**Leire**, cariño, tengo tanto para agradecerte a vos. Gracias por ser mi amiga, la que me saca una sonrisa en mis momentos más "refunfu" y la que me apoya cuando siento que no puedo más. Gracias por ser mi compañera de aventuras, la que me motiva a hacer cosas nuevas, locas, divertidas y que (casi) siempre me dan un máximo cagazo. Gracias por ser mi maestra, y enseñarme a ser más compañero y más cercano con la gente. Gracias por abrirme las puertas de tu casa y por hacerme sentir parte de tu familia! Gracias por muchas otras cosas más, las cuales no se pueden detallar en un libro de carácter académico xD. La verdad, hacer esta tesis sin vos, hubiese sido muy muy difícil o, quizás, me hubiese llevado a la locura. Quizás ahora no sabemos bien cuál es nuestro norte, pero estoy seguro de que vamos hacia allí juntos. Te amo mucho.

## About the Author

Julián Andrés Paganini was born on the 7th of September 1988, in Rosario, Santa Fe, Argentina. In 2017 he obtained his Masters Degree in Biotechnology at the Universidad Nacional de Rosario. For his Master's thesis, he joined the laboratory of Dr. Hugo Gramajo at the Institute of Molecular Biology of Rosario (IBR). Under the supervision of Dr. Ana Arabolaza and Dr. Simon Menendez-Bravo, he engineered multiple strains of *Streptomyces coelicolor* in order to increase the production of triacylglycerols, a key precursor for the synthesis of biofuels. After finishing graduate school, he developed a strong interest in bioinformatics, and in 2018 he started an internship at the laboratory of Dr. Alejandro Viale, also at the IBR. During this internship, he studied the role of insertion sequences (IS) in determining the genome plasticity and pathogenicity of clinical isolates of *Acinetobacter baumannii*. In January 2020, he started his PhD thesis at the University Medical Center of Utrecht, in the Netherlands, working under the supervision of Dr. Anita C. Schürch, Dr. Nienke L. Plantinga and Prof. Rob Willems. During the past three years he has performed the work presented in this thesis.

